Tracking Large-Scale Video Remix in Real-World Events

Lexing Xie, Apostol Natsev, Xuming He, John R. Kender, Matthew Hill, and John R. Smith

***Abstract*—Content sharing networks, such as YouTube, contain traces of both explicit online interactions (such as likes, comments, or subscriptions), as well as latent interactions (such as quoting, or remixing, parts of a video). We propose visual memes, or frequently re-posted short video segments, for detecting and monitoring such latent video interactions at scale. Visual memes are extracted by scalable detection algorithms that we develop, with high accuracy. We further augment visual memes with text, via a statistical model of latent topics. We model content interactions on YouTube with visual memes, defining several measures of influence and building predictive models for meme popularity. Experiments are carried out with over 2 million video shots from more than 40,000 videos on two prominent news events in 2009: the election in Iran and the swine flu epidemic. In these two events, a high percentage of videos contain remixed content, and it is apparent that traditional news media and citizen journalists have different roles in disseminating remixed content. We perform two quantitative evaluations for annotating visual memes and predicting their popularity. The proposed joint statistical model of visual memes and words outperforms an alternative concurrence model, with an average error of 2% for predicting meme volume and 17% for predicting meme lifespan.**

***Index Terms*—Image databases, YouTube, social networks.**

I. Introduction

THE ease of publishing and sharing videos online has led to an unprecedented information explosion [3], outpacing the ability of users to discover and consume such content. This information overload problem is particularly prominent for linear media (such as audio, video, animations), where at-a-glance impressions are hard to develop and often unreliable. While text-based information networks such as Twitter rely on re-tweets [27] and hashtags [34] to identify influential and trending topics, similar capabilities of "quoting" video and tracking video reuse do not exist. On the other hand, video clipping and remixing is an essential part of the participatory culture on sites like YouTube [37]. Moreover, a reliable video-based "quoting" mechanism, video remix tracking, and popularity analysis, could be used in multiple domains, including brand monitoring, event spotting for emergency management, trend prediction, journalistic content selection, or better retrieval.

We propose to use visual memes, or short segments of video frequently remixed and reposted by multiple authors, as a tool for making sense of video "buzz". Video-making and remixing requires a significant effort and time, therefore we consider reposting a video meme as a deeper stamp of approval—or awareness—than simply commenting, rating, or tweeting a video. Example video memes are shown in Fig. 1, represented as static keyframes. We develop a large-scale event monitoring system for video content, using generic text queries as a pre-filter for content collection on a given topic. We apply this system to collect large video datasets over a range of topics on YouTube. We then perform fast and accurate visual meme detection on tens of thousands of videos and millions of video shots. We augment the detected visual memes with relevant text using a statistical topic model, and propose a Cross-Modal Matching (CM^2) method to automatically explain visual memes with corresponding textual words. We design a graph representation for social interactions via visual memes, and then derive graph metrics to quantify content influence and user roles. Furthermore, we use features derived from the video content and from meme interactions to model and predict meme popularity, with an average error of 2% on the volume and 17% on the lifespan prediction.

This work is an expanded version of a conference publication on visual memes [41], based on the same data collection and video remix detection method [25]. The following components are new since [41]: CM^2, a new representation for visual memes and their associated text, new approach and results on predicting meme popularity, and expanded discussion on related work. The overall contributions of this work include:

- We propose *visual memes* as a novel tool to track large-scale video remixing in social media.
- We design a scalable system capable of detecting memes from over a million video snippets in a few hours on a single machine.
- We design and implement a large-scale event-based social video monitoring and content analysis system.
- We design a novel method, CM^2, to explain visual memes with statistical topic models.
- We design a graph model for social interaction via visual memes for characterizing information flow and user roles in content dissemination.

Manuscript received September 18, 2012; revised January 31, 2013 and April 24, 2013; accepted April 29, 2013. Date of publication May 29, 2013; date of current version September 13, 2013. NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program. This work was supported in part by the US Air Force Research Laboratory, under agreement number FA2386-12-1-4041. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory or the U.S. Government. This work was done while L. Xie, A. Natsev, and J. R. Kender were with IBM Research. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Daniel Gatica-Perez.

L. Xie and X. He are with the Australian National University (ANU) and National ICT Australia (NICTA), Canberra 0200, Australia.

A. Natsev is with Google Research, Mountain View, CA 94043 USA.

J. R. Kender is with Columbia University, New York, NY 10027 USA.

M. Hill and J. R. Smith are with the IBM Watson Research Center, Yorktown Heights, NY 10598 USA .

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2013.2264929

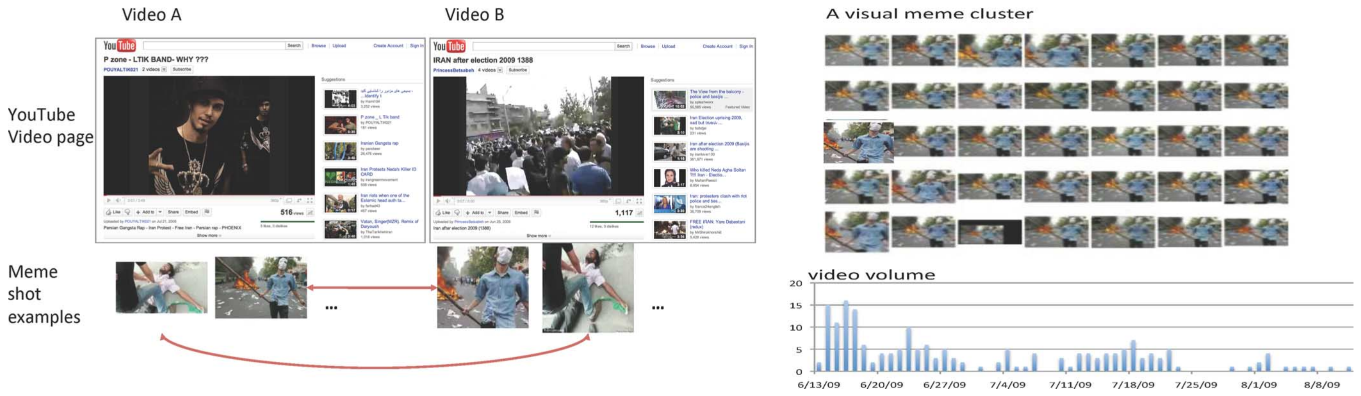

Fig. 1. Visual meme shots and meme clusters. (Left) Two YouTube videos that share multiple different memes. Note that it is impossible to tell from metadata or the YouTube video page that they shared any content, and that the appearance of the remixed shots (bottom row) has large variations. (Right) A sample of other meme keyframes corresponding to one of the meme shots, and the number of videos containing this meme over time —193 videos in total between June 13 and August 11, 2009.

- We conduct empirical analysis on several large-scale event datasets, producing observations about the extent of video remix, the popularity of memes against traditional metrics, and various diffusion patterns.

The rest of this paper is organized as follows. Section II presents a system design for real-world event monitoring on video sharing sites like YouTube; Section III covers the meme detection algorithm; Section IV proposes the CM^2 model for annotating visual memes; Section V explores a graph representation for meme diffusion, Section VI uses graph features to predict content importance; Section VII discusses our experimental setup and results; Section VIII reviews related work; and Section IX concludes with a summary of this work.

II. Visual Remix and Event Monitoring

In this section, we define visual memes as the unit for video remix tracking, and describe our system to monitor video traces from real-world events.

A. Visual Memes and Online Participatory Culture

The word *meme* originally means [1] “an element of a culture or system of behaviour passed from one individual to another by imitation or other non-genetic means”, and for digital artifacts, it refers to “an image, video, piece of text, etc., typically humorous in nature, that is copied and spread rapidly by Internet users, often with slight variations”. The problem of automatically tracking online memes has been recently addressed for text quotes [28] from news and blogs, as well as for edited images [26] from web search. The scope of this work is to study memes in the context of online video sharing, and in particular, on YouTube.

Media researchers observe that users tend to create “curated selections based on what they liked or thought was important” [37], and that remixing (or re-posting video segments) is an important part of the “participatory culture” [13] of YouTube. Intuitively, re-posting is a stronger endorsement requiring much more effort than simply viewing, commenting on, or linking to the video content. A re-posted visual meme is an explicit statement of mutual awareness, or a relevance statement on a subject of mutual interest. Hence, memes can be used to study virality, lifetimes and timeliness, influential originators, and (in)equality of reference.

We define visual memes as frequently reposted video segments or images, and this study on video remix has two operational assumptions. The first is to focus on videos about particular news events. Using existing footage is common practice in the reporting and discussion of news events [30]. The unit of reuse typically consists of one or a few contiguous shots, and the audio track often consist of re-dubbed commentary or music. The second assumption is to restrict remix-tracking to short visual segments and to ignore audio. The remixed shots typically contain minor modifications that include video formatting changes (such as aspect ratio, color, contrast, gamma) and production edits (such as the superimposing text, or adding borders and transition effects). Most of these transformations are well-known as the targets of visual copy detection benchmarks [33].

Using the above definition and assumptions, this work proposes tools for detecting visual remixes, and for quantifying their prevalence. The observations specific to news topics do not readily generalize to the entire YouTube, or to video genres designed for creativity and self-expression, such as video blogs. In the rest of this paper, *meme* refers both to individual instances, visualized as representative icons (as in Fig. 1 Left), and to the entire equivalence class of re-posted near-duplicate video segments, visualized as clusters of keyframes (as in Fig. 1 Right).

B. Monitoring Events on YouTube

We use text queries to pre-filter content, and make the scale of monitoring feasible [3]. We use a number of generic, time-insensitive text queries as content pre-filters. The queries are manually designed to capture the topic theme, as well as the generally understood cause, phenomena, and consequences of the topic. For example, our queries about the “swine flu” epidemic consist of *swine flu, H1N1, H1N1 travel advisory, swine flu vaccination.*[1] We aim to create queries covering the key invariant aspects of a topic, but automatic time-varying query expansion is open for future work. We use the YouTube API to extract video entries for each query, sorted by relevance and recency, respectively. The API will return up to 1000 entries per query, so

[1] The full set of queries is available on the accompanying webpage [2].

varying the ranking criteria helps to increase content coverage and diversity. Then, for each unique video, we segment it into shots using thresholded color histogram differences. For each shot we randomly select and extract a frame as the keyframe, and extract visual features from each keyframe. We process the metadata associated with each video, and extract information such as author, publish date, view counts, and free-text title and descriptions. We clean the free-text metadata using stop word removal and morphological normalization. The volume of retrieved and memes are telling indicators of event evolution in the real world, a few example trends can be found in our recent paper [41] and on the project webpage [2].

III. Scalable Visual Meme Detection

Detecting visual memes in a large video collection is a nontrivial problem. There are two main challenges. First, remixing online video segments changes their visual appearance, adding noise as the video is edited and re-compressed (Section II). Second, finding all pairs of near-duplicates by matching all N shots against each other has a complexity of $O(N^2)$, which is infeasible for any reasonably large collections.

Our solution to keyframe matching has three parts, each contributing to the robustness of the match. Here a keyframe is representative of a video shot, segmented using temporal feature differences. We first pre-process the frame by removing trivial (e.g., blank) matches, detecting and removing internal borders; normalizing the aspect ratio; de-noising with median filters; and applying contrast-limited histogram equalization to correct for contrast and gamma differences. We then extract the *color correlogram* [23] feature for each frame to capture the local spatial correlation of pairs of colors. The color correlogram is designed to tolerate moderate changes in appearance and shape that are largely color-preserving, e.g., viewpoint changes, camera zoom, noise, compression, and to a smaller degree, shifts, crops, and aspect ratio changes. We also use a "cross"-layout that extracts the descriptor only from horizontal and vertical central image stripes, thereby emphasizing the center portion of the image and improving robustness with respect to text and logo overlay, borders, crops, and shifts. We extract an auto correlogram in a 166-dimensional perceptually quantized HSV color space, resulting in a 332-dimensional feature. Finally, the system uses a frame-adaptive threshold on pairwise frame similarity, normalized across frame complexity and the corresponding feature entropy. This threshold is tuned on a training set. Detailed comparison of each technique can be found in a related paper [33].

Our solution to the complexity challenge is to use an indexing scheme for fast approximate nearest neighbor (ANN) look-up. We use the FLANN Library [32] to automatically select the best indexing structure and its appropriate parameters for a given dataset. Our frame features have over 300 dimensions, and we empirically found that setting the number of nearest-neighbor candidate nodes m to $\sqrt{N}$ can approximate k-NN results with approximately 0.95 precision. In running in $O(N\sqrt{N})$ time, it achieves two to three decimal orders of magnitude speed-up over exact nearest neighbor search. Furthermore, each FLANN query results in an incomplete set of near-duplicate pairs so we perform transitive closure on the neighbor relationship to find equivalence classes of near-duplicate sets. We use an efficient set union-find algorithm [19] that runs in amortized time of $O(E)$, where E is the number of matched pairs, which is again $O(N\sqrt{N})$.

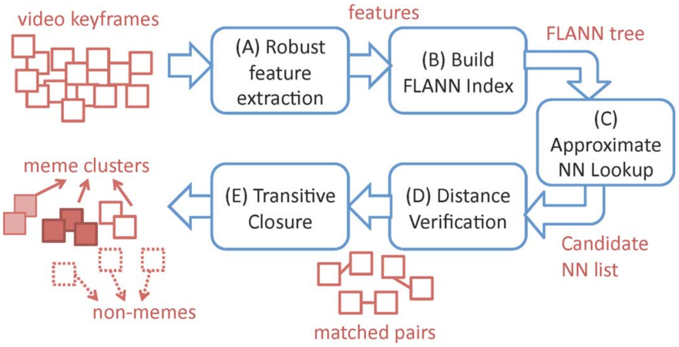

Fig. 2. Flow diagram for visual meme detection method.

This process for detecting video memes is outlined in Fig. 2. The input to this system is a set of video frames, and the output splits this set into two parts. The first part consists of a number of meme clusters, where frames in the same cluster are considered near-duplicates with each other. The second part consists of the rest of the frames that are not considered near-duplicates with any other frame. Blocks A and D address the robust matching challenge using correlogram features and frame-adaptive thresholding, and blocks B, C and E address the scalability challenge using approximate nearest-neighbor (ANN) indexing. A few examples of identified near-duplicate sets are shown in Fig. 1. Visual meme detection performance is evaluated in Section VII.A.

Our design choices for the visual meme detection system aim to find a favorable combination of accuracy and speed feasible to implement on a single machine. Note that local image points or video sequences [39] tend to be accurate in each query, but is not easy to scale to N^2 matches. We found that a single video shot is a suitable unit to capture community video remixing, and that matching by video keyframe is amenable to building fast indexing structures. The ANN indexing scheme we adopt scales to several million video shots. On collections consisting of tens of millions to billions of video shots, we expect that the computation infrastructure will need to change, such as using a data center to implement a massively distributed tree structure [29] and/or hybrid tree-hashing techniques.

IV. Topic Representation of Memes

Visual memes are iconic representations for an event, it will be desirable to augment its image-only representation with textual explanations. It is easy to see that the title and text descriptions associated with many online videos can be used for this purpose, despite the noisy and imprecise nature of the textual content. We propose to build a latent topic model over both the visual memes and available text descriptions, in order to derive a concise representation of videos using memes, and to facilitate applications such as annotation and retrieval.

Our model treats each video as a document, in which visual memes are "visual words" and the annotations are text words. By building a latent topic space for video document collections, we embed the high-dimensional bag-of-words into a more concise and semantically meaningful topic space. Specifically, we learn a set of topics $z = 1, \ldots, K$ on the multimedia document collection $\mathcal{D} = \{d_m, m = 1, \ldots, M\}$ using latent Dirichlet Allocation (LDA) [10]. LDA models each document as a mixture of topics with a document-dependent Dirichlet prior, each topic drawn from the resulting multi-nomial, and each word drawn from a topic-dependent multi-nomial distribution. Our LDA model combines two types of words, i.e., visual memes and text words, into a single vocabulary $\mathcal{V} = \{\mathcal{V}_v, \mathcal{V}_t\}$, and estimates the conditional distribution Φ of words given topics from a video collection. Mathematically, each video d_m is represented as a bag of words, $\mathbf{w}_m = \{w_i\}_{i=1}^{N_m}$, is modeled as follows,

$$P(d_m|\alpha) = \int_\theta \prod_{w_i \in d_m} \left[\sum_{z_i} P(w_i|z_i, \Phi) P(z_i|\theta)\right] P(\theta|\alpha) d\theta \tag{1}$$

where z_i is the topic indicator variable for word w_i, θ is the latent topic distribution, and α is the hyperparameter controlling the topic prior at the corpus level.

Given the topic model, we can project a set of visual words (or text tags) into the learned topic space by computing the posterior of the topic distribution θ conditioned on those words. Let the observed words are $\mathbf{w}_o$, we map $\mathbf{w}_o$ to the mode $\hat{\theta}$ of the topic posterior:

$$\hat{\theta} = \arg\max_\theta \prod_{w_i \in \mathbf{w}_o} \sum_{z_i} P(w_i|z_i) P(z_i|\theta, \Phi) P(\theta|\alpha) \tag{2}$$

where the parameters α, Φ are estimated from training data. The inference in model learning and posterior calculation are conducted with variational EM (e.g., see [10] for details).

A. *Cross-Modal Matching* (CM^2) *With Topics*

In social media, some words and names may be unseen in previous events (e.g., *entekhabat*, "election" in Persian), and iconic visual memes may appear without clear context of emergence. For a better understanding of these novel events, a particular useful step is to build association between different modalities, such as texts and visual memes. We pose this as a cross-modal matching problem, and aim to estimate how well a textual or visual word (candidate result w_r) can explain another set of words (query $\mathbf{w}_q = \{w_{q_n}\}$). This is achieved by estimating the conditional probability of seeing w_r given that $\mathbf{w}_q$ is in the document, i.e., $p(w_r|\mathbf{w}_q, D)$. We call this estimation process Cross-Modal Matching (CM^2), and propose its application for content annotation and retrieval.

A derivation sketch for CM^2 is as follows, under the context of document collection D and the topic model $\{\alpha, \Phi\}$. We consider modeling the conditional word distribution through topic representation. Without loss of generality, we assume the query consists of visual memes and predict the probability of each text tag. The first step of our method is to compute the topic posteriors of the document collection $D = \{d_m\}$ given the query modality. Let $\mathbf{w}_o$ be the observed visual memes in each document d_m, we estimate the topic posterior mode $\hat{\theta}_m$ from (2). Thus the whole document collection can be represented as $\Theta = \{\theta_m\}$.

Given a new query $\mathbf{w}_q$, we also embed it into the topic space by computing its topic posterior mode:

$$\theta_q = \arg\max_\theta \prod_{w_i \in \mathbf{w}_q} \sum_{z_i} P(w_i|z_i) P(z_i \theta, \Phi) P(\theta|\alpha)$$

Intuitively, we want to use "similar" videos in the topic space to predict the text tag probability. Formally, the conditional probability of a text tag w_r is estimated by a non-parametric voting scheme as follows,

$$P(w_r|\mathbf{w}_q, D) \propto \sum_m \left(\sum_{w_i \in d_m} [w_i = w_r]\right) e^{-(\theta_q - \theta_m)^2/\sigma_\theta} \tag{3}$$

where σ_θ controls the similarity of topic vectors and is set to the median of the training data.

A baseline method based on word co-occurrence can estimate the conditional probability with co-ocurrence counting:

$$P(w_r|\mathbf{w}_q, D) \propto \sum_m \left(\sum_{w_i \in d_m, q_n} [w_i = w_r] \wedge [w_{q_n} \in \mathbf{w}_q]\right) \tag{4}$$

Examining the estimation (3)–(4), we note that CM^2 can be interpreted as a soft co-occurrence measure for $(w_r, \mathbf{w}_q)$ over the entire document collection with the topic model. In a sense, co-occurrence counting is a special case, where the counts are weighted by the number of documents in which $\mathbf{w}_q$ appeared.

B. CM^2 *Applications*

CM^2 has several applications depending on the choice of w_q and w_r. Such as (1) Visual Meme/video annotation—We use visual memes as queries, $\mathbf{w}_q \subset \mathcal{V}_v$, and return the top entries of $w_r \in \mathcal{V}_t$, sorted by $p(w_r|\mathbf{w}_q, D)$. The motivation of this task for event monitoring is that the keywords are often specialized subjective, semantic, and non-visual, e.g.,*freedom*. (2) Keyword illustration—We can illustrate a keyword (e.g., *H1N1*) with a set of most-related images. We take $w_q \in \mathcal{V}_t$, and yield the top entries of $w_r \in \mathcal{V}_v$, sorted by $p(w_r|w_q, D)$. In this paper, we focus on application (1) and leave the others for future exploration.

V. Graphs on Visual Memes

Visual memes can be seen as implicit *links* between videos and their creators that share the same unit of visual expression. We construct graph representations for visual memes and users who create them. This gives us a novel way to quantify influence and the importance of content and users in this video-centric information network.

Denote a video (or multimedia document) as d_m in event collection $\mathcal{D}$, with $m = 1, \ldots, N$. Each video is authored (i.e., uploaded) by author $a(d_m)$ at time $t(d_m)$, with $a(d_m)$ taking its value from a set of authors $\mathcal{A} = \{a_r, r = 1, \ldots, R\}$. Each video document d_m contains a collection of memes, $\{v_1, v_2, \ldots, v_{K_m}\}$ from a meme dictionary $\mathcal{V}$. In this network model, each meme induces a time-sensitive edge e_{mj} with creation time $t(e_{mj})$, where m, j are over video documents or authors.

A. Meme Video Graph

We define the video graph $G = \{\mathcal{D}, \mathcal{E}_G\}$, with nodes $d \in \mathcal{D}$. There is a directed edge $e_{mj} \in \mathcal{E}_G$, if documents d_m and d_j share at least one visual meme, and if d_m precedes d_j in time, with $t(d_m) < t(d_j)$. The presence of e_{mj} represents a probability that d_j was derived from d_m, even though there is no conclusive evidence within the video collection alone whether or not this is true. We denote the number of shared visual memes as $\nu_{mj} = |\, d_m \cap d_j \,|$, and the time elapsed between the posting time of the two videos as $\triangle t_{jm} = t(d_j) - t(d_m)$.

We use two recipes for computing the edge weight ω_{mj}. Equation (5) uses a weight proportional to the number of common memes ν_{mj}, and (6) scales this weight by a power-law memory factor related to the time difference $\triangle t_{jm}$. The first model is insensitive to $\triangle t_{jm}$, so it can accommodate the resurgence of popular memes, as seen in textual memes [28]. The power law decay comes from known behaviors on YouTube [17], and it also agrees with our observations on the recency of the content returned by the YouTube search API.

$$\omega^*_{mj} = \nu_{mj}(m, j) \in \mathcal{E}_G \quad (5)$$

$$\omega'_{mj} = \nu_{mj} \triangle t_{jm}^{-\eta} \quad (6)$$

We estimate the exponent η to be 0.7654, by fitting an exponential curve to the video age versus volume to a subset of our data, over ten different topics retrieved over 24 hours of time.

B. Meme Author Graph

We define an author graph $H = \{\mathcal{A}, \mathcal{E}_H\}$, with each author $a \in \mathcal{A}$ as nodes. There is an undirected edge $e_{rs} \in \mathcal{E}_H$, if authors a_r and a_j share at least one visual meme in any video that they upload in the event collection.

We compute the edge weights θ_{rs} on edge e_{rs} as the aggregation of the weights on all the edges in the video graph G connecting documents authored by a_r and a_s.

$$\theta_{rs} = \Sigma_{\{i, a(d_m)=a_r\}} \Sigma_{\{j, a(d_j)=a_s\}} \omega_{mj} \quad (7)$$

with $r, s \in \mathcal{A}, m, j \in \mathcal{D}$. We adopt two simplifying assumptions in this definition. The set of edges $\mathcal{E}_H$ are bidirectional, as authors often repost memes from each other at different times. The edge weights are cumulative over time, because in our datasets most authors post no more than a handful of videos (Fig. 7), and there is rarely enough data to estimate instantaneous activities.

C. Meme Influence Indices

We define three indices based on meme graphs, which captures the influence on information diffusion among memes, and in turn quantifies the impact of content and of authors within the video sharing information network.

First, for each visual meme v, we extract from the event collection $\mathcal{D}$ the subcollection containing all videos that have at least one shot matching meme v, denoted as $\mathcal{D}_v = \{d_j \in \mathcal{D}, \text{ s.t. } v \in d_j\}$. We use $\mathcal{D}_v$ to extract the corresponding video document subgraph G_v and its edges, setting all edge weights ν in G_v to 1 since only a single meme is involved. We compute the in-degree and out-degree of every video d_m in $\mathcal{D}_v$ as the number of videos preceding and following d_m in time:

$$\begin{aligned} \zeta^{in}_{m,v} &= \Sigma_j I\{d_m, d_j \in \mathcal{D}_v, \quad t(d_j) < t(d_m)\} \\ \zeta^{out}_{m,v} &= \Sigma_j I\{d_m, d_j \in \mathcal{D}_v, \quad t(d_j) > t(d_m)\} \end{aligned} \quad (8)$$

where $I\{\cdot\}$ is the indicator function that takes a value of 1 when its argument is true, and 0 otherwise. Intuitively, ζ^{in}_m is the number of videos with meme v that precede video d_m (potential sources), and ζ^{out}_m is the number of videos that succeed meme v after video d_m (potential followers).

The video influence index χ_m is defined for each video document d_m as the smoothed ratio of its out-degree over its in-degree, aggregated over all meme subgraphs G_v (9); the smoothing factor 1 in the denominator accounts for d_m itself). The author influence index $\hat{\chi}_r$ is obtained by aggregating χ_m over all videos from author a_r (10). The normalized author influence index $\bar{\chi}_r$ is its un-normalized counterpart $\hat{\chi}_r$ divided by the number of videos an author posted, which can be interpreted as the *average* influence of all videos for this author.

$$\chi_m = \Sigma_v \frac{\zeta^{out}_{m,v}}{1 + \zeta^{in}_{m,v}} \quad (9)$$

$$\hat{\chi}_r = \Sigma_{\{i, a(d_m)=a_r\}} \chi_m, \quad (10)$$

$$\bar{\chi}_r = \frac{\hat{\chi}_r}{\Sigma_m I\{a(d_m) = a_r\}}$$

The influence indexes above captures two aspects in meme diffusion: the volume of memes, and how "early" a video or an author is in the diffusion chain. The first aspect is similar to the *reweet* and *mention* measures recently reported on Twitter [14]. The timing aspect in diffusion is designed to capture different roles that users play on YouTube, These roles can be intuitively understood as information *connectors* and *mavens* [21]. Here *connectors* refer to people who come "... with a special gift for bringing the world together,... [an] ability to span many different worlds", and *mavens* are "people we rely upon to connect us with new information, ... [those who start] word-of-mouth epidemics".

The meme video and author graphs are used to generate features that describe node centrality and meme diffusion history, which are in turn used to predict importance of visual memes in Section VI.

VI. Predicting Meme Importance

One long-standing problem in social media is on predicting the popularity of social memes [21]. Studies on social meme adoption and popularity has focused on URLs [6], hashtags [44], or view counts on YouTube [11], [38]. This work investigates whether or not visual meme popularity is predictable with knowledge of both the network and content.

Popularity, or importance on social media is inherently multidimensional, due to the rich interaction and information diffusion modes. For YouTube, it can be the number of times that a video is viewed [38], the number of likes or favorites that a video has received. While these commonly-used metrics focus on the entire video, not a given meme, we focus on two targets that are inherent to visual memes: the number of times that a video meme is reposted by other YouTube users (denoted as *volume*), or by the lifespan (in days) of a meme (*life*).

We build a meme prediction model using features related to its early dynamics, the network around its authors, and the visual and textual content. For each visual meme v that first appeared at time $t(v)$ (called *onset* time), we compute features on the meme- and author- sub-graphs up to $t_1 = t(v) + \Delta t$, by including video nodes that appeared before t_1. Δt is set to one day in this work, to capture meme early dynamics, similar to what has been used for view-count prediction [38].

There are five types of features in total. For features aggregated over multiple authors, we take the maximum, average, median, and standard deviation among the group of authors who have posted or reposted the meme by t_1. (1) *volume-d1*, 1 dimension—the volume of memes up to t_1. (2) *connectivity*, 28 dimensions—static network features of author productivity and connectivity. We use the *total number of videos* that the author has uploaded to capture author productivity. An author's connectivity include three metrics computed over the author and video graphs, respectively of up to time t_1: *degree centrality* is the fraction of other nodes that a node is directly connected to; *closeness centrality* is the inverse of average path length to all other nodes; and *betweenness centrality* is the fraction of all-pairs shortest paths that pass through a node [12]. (3) *influence*, 16 dimensions—dynamic features of author diffusion influence. These include the meme influence indices $\hat{\chi}_r$ and $\bar{\chi}_r$ in (10), as well as the aggregated in-degree and out-degree for each author. (4) *txt*, 2000 dimensions—the bag-of-word vector for each meme v is the average count of each word over all videos containing v within the first day; *vmeme*—bag of visual meme vectors compiled in the same way as *txt*, with 2000 dimensions for the *Iran3* and 400 dimensions for *Swineflu*, respectively; *topic*, 50-dimensional vector is the posterior probability of each topic given v inferred through the topic model in Section IV.A.

We use Support Vector Regression (SVR) [16] to predict meme volume and lifespan on a log-scale, using each, and the combination the features above.

VII. Experiments and Observations

This section will first present our experimental setup, and then discuss our main results as both quantitative observations and as quantitative evaluations. The former include observations on video popularity versus remix probability (Section VII.B), and on apparent temporal diffusion patterns (Section VII.D); the latter include visual meme detection (Section VII.A), cross-modal annotation (Section VII.C), and popularity prediction (Section VII.E).

Using the targeted-querying and collection procedures described in Section II.B, we downloaded video entries from about a few dozen topics from May 2009 to March 2010. We used the following four sets for evaluation, which had enough volume and change over time to report results, summarized in Table I. The SwineFlu set is about the H1N1 flu epidemic. The Iran3 set is about Iranian domestic politics and related international events during the 3-month period of summer 2009. The Housing set is about the housing market in the 2008–09 economic crisis; a subset of these videos were manually annotated and used to validate and tune the visual meme detection algorithms.

TABLE I
Summary of YouTube Event Data Sets

Topic	#Videos	#Authors	#Shots	Upload time
SwineFlu	31,488	10,804	1,202,479	04/09~03/10
Iran3	23,049	4,681	1,255,062	08/07~08/09
Housing	2,446	654	71,872	08/07~08/09

We perform visual meme detection as described in Section III. We additionally filter the meme clusters identified by the detection system, by removing singletons belonging to a single video or a single author. We process words in the title and description of each video by morphological normalization with a dictionary [4], we preserve all out-of-vocabulary words, these include foreign words and proper names (e.g., *mousavi*), abbreviations (*H1N1*), or mis-spellings. We rank the words by tf-idf across all topics, and take the top few thousand for topic models, tagging, and popularity prediction. The prototype system is implemented in C++, Python, and MATLAB, and it can be deployed on one workstation requiring less than 8 GB memory.

A. Meme Detection Performance

We evaluate the visual meme detection method described in Section III using a test set created from the Housing dataset. This is done by one annotator who is not aware of the detection algorithm, and she was instructed to find visually *very similar* keyframes that are likely to be taken at the same scene. Specifically, she start from seed sets created from multiple runs of k-means clustering with a tight cluster radius threshold, and top 50 returns based on color feature similarity using multiple random keyframes as the query. The annotator manually goes through these results to explicitly mark the clusters as correct and incorrect near-duplicates, and the top returns as duplicates with the query or not. This annotation protocol specifically includes many pairs that are being confused by the clustering and feature-similarity retrieval steps. The resulting data set contains $\sim 15,000$ near-duplicate keyframe pairs and $\sim 25,000$ non-duplicate keyframe pairs.

We compute the near-duplicate equivalence classes as described in Section III, and calculate precision (P) and recall (R) on the labeled pairs. The results are shown on Fig. 3 for varying values of the threshold parameter τ. We note that the performance is generally quite high with $P > 95\%$. There are several possible operating points, such as $P = 99.7\%, R = 73.5\%$ for low false alarm; or $P = 98.2\%, R = 80.1\%$ that produces the maximum F1 score of 0.88 (defined as $\frac{2PR}{P+R}$); or $P = 96.6\%, R = 80.7\%$ for the highest recall. For the rest of our analysis, we use the last, high-recall, point with $\tau = 11.5$. On the Iran3 set of over 1 million shots, feature extraction takes around 7 hours on a quad-core CPU, and indexing and querying with FLANN takes 5 to 6 CPU hours.

B. Content Views and Re-Posting Probability

In our video collections, the behavior of remixing and reposting is quite dominant. Over 58% of the videos collected for the Iran3 topic contain at least one visual meme, and 70% of the authors participate in remixing; likewise, 32% and 45% respectively for SwineFlu, as shown in Fig. 4(a). These statistics

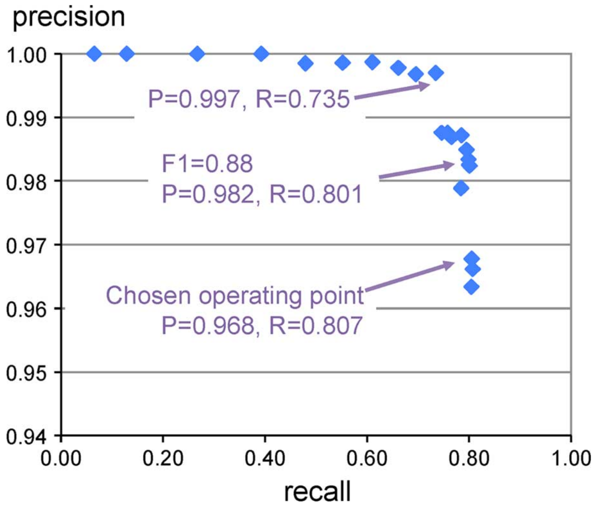

Fig. 3. Performance of visual meme detection method on Housing Dataset.

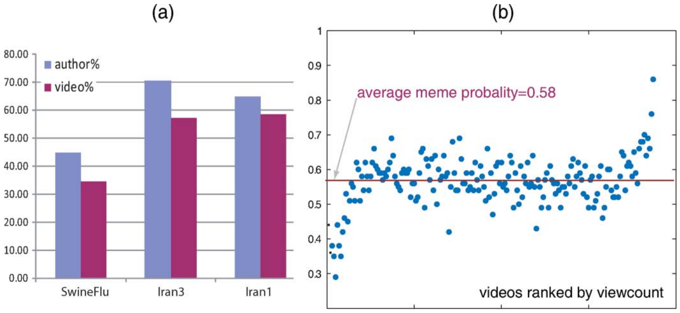

Fig. 4. Video reposting probabilities. (a) Fraction of visual memes. (b) Video views vs. meme probability on Iran3 set.

suggest that, for content covering trending events, a significant fraction of the authors re-mix existing sources.

Fig. 4 shows empirical estimates of a video containing at least one meme in the Iran3 set, binned by video view count (descending from left to right). We observe that the 4 most viewed videos have no memes and have nothing to do with the topic, and likewise for 7 of the first 10. One has to get beyond the first 1,600 most popular videos before the likelihood of having near-duplicates passes the average for the dataset, at about 0.58 (see Fig. 4(b)). The main reason for this mismatch is likely that some of the queries were not well targeted (e.g., “Iran”), returning generally popular videos that were off-topic for the specific event. Popular videos often come from entertainment-oriented verticals (e.g., music, gaming), which bear lesser value for re-interpretation as compared to news events. This may have led to skewing of the results for the popular videos returned for each topic. In the Iran topic for example, the video with the highest view-count is a music video irrelevant to Iranian politics.

We observe considerable similarity in the frequency distribution of visual memes and words. Fig. 5 is a scatter plot of textual words and visual memes ranked by descending frequency in log-log scale. Performing a regression fit, we obtain the following Zipf’s power law distributions:

$$f(w_t) \propto r^{1.102}; f(w_v) \propto r^{1.959}$$

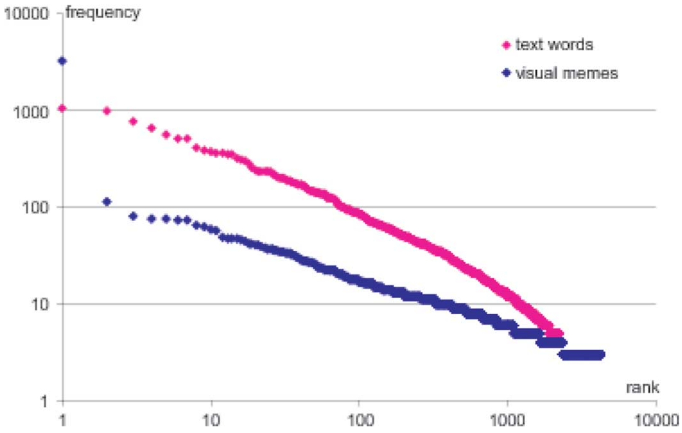

Fig. 5. Rank vs. frequency for words and visual memes.

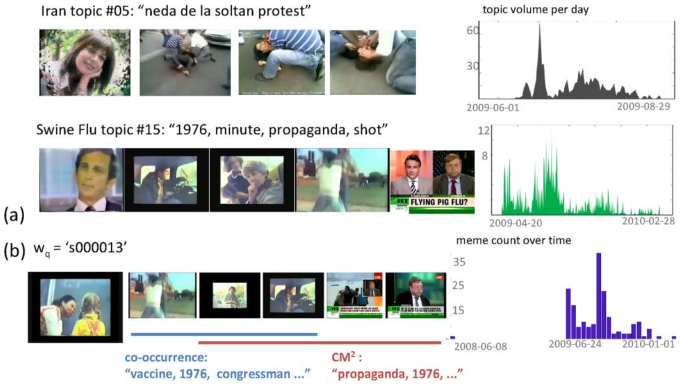

Fig. 6. (a) Topic model example. (b) Retrieval results using CM2. Please view in color and magnified for optimal readability.

The exponent s for words in the title and description is close to that of English words (~ 1.0). For visual memes $s = 1.959$, suggesting that the diversity of visual memes is less than that of words at the lower-frequency end. This suggest that it makes sense to model visual words appearances in a similar way as those with textual words.

C. Multimodal Topics and CM^2

We learn topic models on a joint vocabulary of words and memes. For words, we adopt a tf-idf re-weighting scheme [31] across more than two dozen topics monitored around the same time, this is to suppress very popular words and yet not overly favor rare words. The visual meme vocabulary is constructed using a threshold on its frequency. In the following experiments, we choose 12000 visual memes for the Iran3 and 4000 visual memes for SwineFlu collection, and 2000 text words for both collections.

We set the number of topics to be 50 by validation, and use the term-topic probabilities $p(w|z)$ to label a topic, using both text words and visual memes. Fig. 6(a) shows two example topics over the collections Iran3 and Swineflu, respectively. Topic $\#5$ contains the keywords “neda, soltan, protest,...”, the images capturing her tragic murder and her portrait that was published later. The topic volume over time clearly showed the onset and peak of the event (just after June 20th, 2009), and we verified

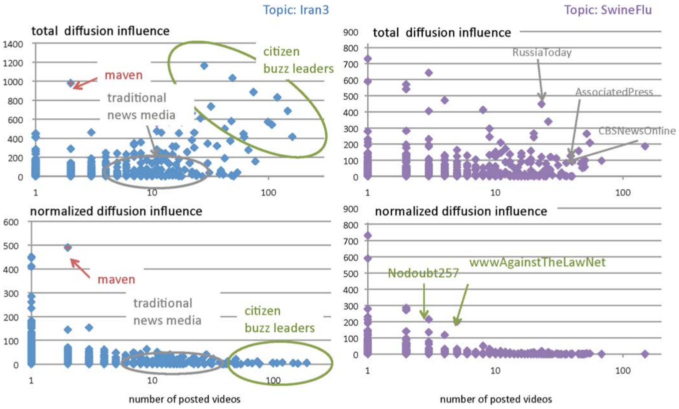

Fig. 7. Meme influence indices vs author productivity on topic Iran3 (Left) and SwineFlu (Right); detailed discussions in Section VII.D. We recommend viewing in color and with magnification.

TABLE II
COMPARISON ON THE LOG-PROBABILITY OF TAGS

Dataset	Co-occurrence	CM^2
Iran3	-6.08 ± 0.06	-5.65 ± 0.05
SwineFlu	-6.59 ± 0.03	-6.54 ± 0.04

from the narrative on Wikipedia that this event also influenced subsequent protest activities in July, corresponding to another peak in meme volume.

We examine the CM2 model for video tagging in context. Here we consider using the visual memes of each video as the query and retrieve the tagging words using scores computed with (3). We also implement the baseline in (4) and look at the memes in comparison with those retrieved by top co-occurrence. We carried out five-fold cross-validation, and report the average performance based on the average log likelihood [7] of the existing tags. We did not use a precision or ranking metric, as tags associated with each video are sparse and positive-only, and many informative tags are missing in the data. Table II shows that the average log likelihood is significantly improved on both datasets, this demonstrates the advantage of the topic-based representation.

Fig. 6(b) shows example retrieval result of using one of the memes in the 1976 video as query. We can see that the co-occurrence model returns *1976, vaccine, congressman,* ... which are all spoken or used as description in the original 1976 government propaganda video, while CM2 returns *1976, propaganda*, which was apparently from the topic above. Comparing the images, we can also see that the top memes returned by the co-occurrence model are all from the same video, since the parody is mostly posted by itself, with little or no remix, while CM2 also retrieves two memes relating to modern-day vaccine discussions in the news media, providing relevant context.

The rightmost column in Fig. 6 shows the temporal evolution of a meme (Fig. 6(b)) and two topics (Fig. 6(a)). We can see the source of the 2008 propaganda video in the meme evolution, it also reveals that there are multiple waves of remix and re-posting around the same theme. The topic evolution, on the other hand, segments out sub-events from the broader unfolding of many themes—with Iran topic #5 representing the murder of Neda and its subsequent influence, and Swine Flu #15 closely correlated to the public discussion on vaccines.

D. Influence Index of Meme Authors

We compute the diffusion index for authors according to (10). Fig. 7 contains scatter plots of the author influence indices on the y-axis, versus "author productivity", (number of videos produced) on the x-axis. For both the Iran3 topic and the SwineFlu topic, we plot the total diffusion influence $\hat{\chi}_r$ and the normalized diffusion influence $\bar{\chi}_r$.

In the Iran3 topic we can see two distinct types of contributors. We call the first contributor type *maven* [21] (marked in red), who are post only a few videos, which tend to be massively remixed and reposted. This particular maven was among the first to post the murder of Neda Soltan, and one other instance of student murder on the street. The former post become the icon of the entire event timeline. We call the second contributor type information *connectors* [21] (circled in green), who tend to produce a large number of videos, and who have high total influence factor, yet has lower influence per video. They aggregate notable content, and serve the role of bringing this content to a broader audience. We examined the YouTube channel pages for a few authors in this group, and they seem to be voluntary political activists with screennames like "iranlover100"; and we can dub them "citizen buzz leaders". Some of their videos are slide shows of iconic images. Note that traditional news media, such as AljezeeraEnglish, AssociatedPress, and so on (circled in gray) have rather low influence metric for this topic, partially because the Iran government severely limited international media participation in the event; most of the event buzz was driven by citizens.

However, the SwineFlu collection behaves differently in its influence index scatterplots. We can see a number of *connectors* on the upper right hand side of the total diffusion scatter. But it turns out that they are the traditional media (a few marked in gray), most of which have a large number (> 40) of videos with memes. The few *mavens* in this topic (marked with green text) are less active than in the Iran topic, and notably they all reposted the identical old video containing government health propaganda for the previous outbreak of swine flu in 1976. These observations suggest that it is the traditional new media who seem to have driven most content on this topic, and that serendipitous discovery of novel content does also exist, but has less diversity.

These visualizations can serve as a tool to characterize influential users in different events. We can find user groups serving as *mavens*, or early "information specialists" [21], and *connectors*, who "brings the rest ... together", and henceforth observe different information dissemination patterns in different events.

E. Meme Popularity Prediction Results

We predict the lifespan of memes as described in Section VI. We prune memes that appear less than 4 times, the remaining memes are randomly split in half for training/testing. The Iran3 dataset has 4081 memes in each of the train/test split, the SwineFlu set has 398. Different features are filtered by a low correlation threshold (0.03) and then concatenated to form a joint feature space. We train support vector regressor [16] by searching over hyperparameters C and several kernel types—linear, polynomial, and radial basis function of different width. We use three different metrics to quantify regression performance: mean-square-error (`mse`), Pearson correlation

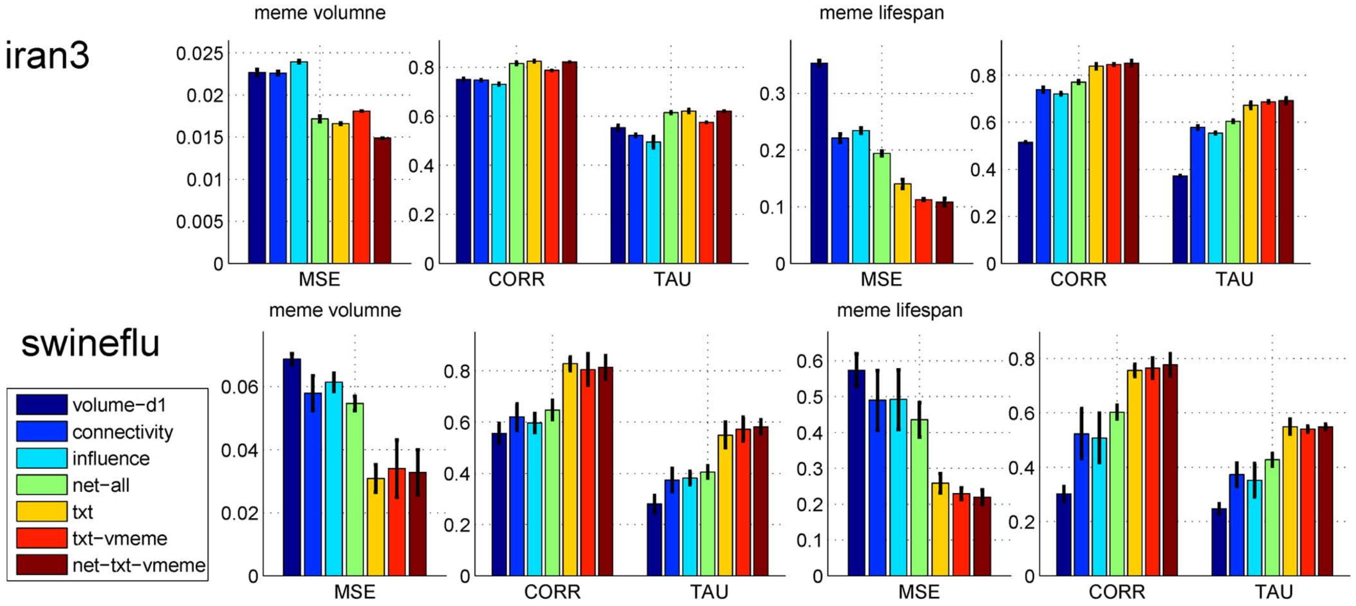

Fig. 8. Meme popularity prediction performance using various network and content features (best viewed in color). Top: Iran3 dataset; bottom: Swineflu dataset. Prediction targets: meme volume (# of videos, left) and lifespan (days, right); performance metrics: M.SE (smaller is better) pearson correlation and Kendall's tau (larger is better). See Section VII.E for discussions on various features.

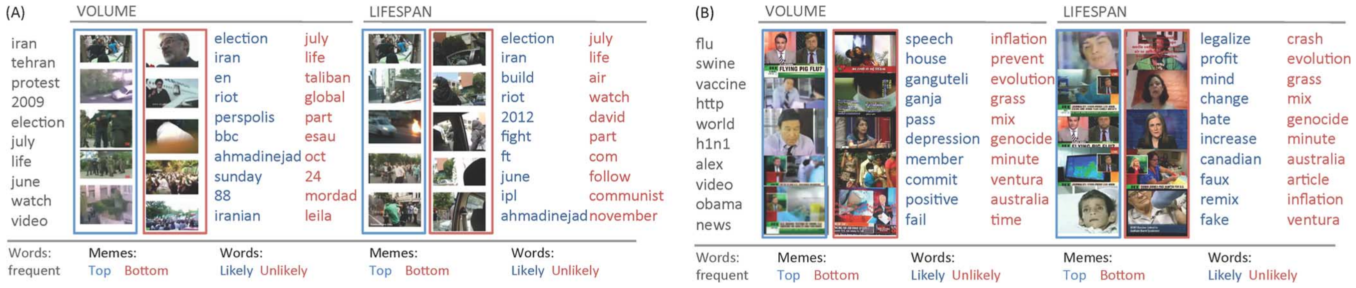

Fig. 9. Illustration of meme popularity prediction results (best viewed in color). (A) Iran3 dataset; (B) SwineFlu dataset. Top popular (blue) and unpopular (red) memes by prediction score; and most likely (blue) and unlikely (red) words by correlation.

(`corr`), Kendall's tau `tau` [24]. Each regressor is trained with five different random splits of train/test data, the average performance with their standard deviation (as error wisks) is shown in Fig. 8.

We can see that meme graph features (*connectivity* and *influence*) both out-perform the baseline feature *volume-d1*. Note that *volume-d1* is the conceptual equivalent of the early view-count features that Szabo and Huberman [38] used to predict long-term view-count on YouTube. Combining these three types of features (*net-all*) further improves prediction performance, and text keyword feature (*txt*) is single strongest predictor. The presence and absence of other visual memes is not stronger than text (*txt+vmeme*), while all of network, words and meme features has the strongest combined performance (*net+txt+vmeme*). The Iran3 dataset, with significantly more data to learn from, has better and more stable results than SwineFlu. From the average MSE, the predictions for meme volume on Iran3 is within 1.7% ($\sqrt{10^{.015}}$) and 16.1% ($\sqrt{10^{.13}}$) for meme lifespan. In Fig. 9 we examine the top- and bottom-ranked visual meme with *net+txt* feature, showing that the top memes are intuitively on-topic, while most of the low-ranked memes have no clear connection to the topic. Fig. 9 also shows the positively and negatively correlated words to each of the prediction target. We can see that they are notably different from frequently-occurring words in either collection. Indicative words include those that indicate trustable authors (*bbc*), particular sub-events (*riot*), or video genre that engender participation (*remix*). On the other hand, certain frequent words such as *watch, video*, or *swine flue, h1n1* are shown to be non-informative.

VIII. Related Work

This work is related to several active research areas in understanding social media and analyzing multimedia content.

Memes have been studied in online information networks of modality other than videos. Retweeting on micro-blogs is a common example [27], where users often quote the original text message verbatim, having little freedom for remixing and context changes within the 140 character limit. MemeTracker [28] detects quoted phrases among millions of blogs and news posts. Kennedy and Chang [26] detects edited images from web search results. Several measurement studies tracked explicit social interactions around online videos. Cha *et al.* [15] characterized content category distribution and exact duplicates of popular videos. Subsequent studies on YouTube include tracking video response actions [8] using metadata, and modeling user comments to determine interesting conversations [18], The audiovisual content of online videos are used to analyze individual social behavior such as personality [9], or used to generate content summaries [22]. Schmitz *et al.* [36] showed that the frequency remix for film content can be used as an implicit video quality indicator. Detecting visual memes on a large-scale and using them to infer implicit social interaction has not been done before.

Our method for detecting visual memes builds upon those for tracking near-duplicates in images and video. Recent foci in near-duplicate detection include speeding up detection on image sequence, frame, or local image points [40], exploring the effect of factors other than visual features [33], and scaling out to web-scale computations using large compute clusters [29]. Compared to copy detection, our work tracks mutual remixes in a large collection rather than matching one query video with reference database. Our method is similar to existing approaches in feature choice, and using approximate nearest-neighbor indexing enables scaling to more than 1 million shots. Topic models [7] is among the popular techniques for the joint modeling of images and text. Our new insight on the CM^2 models is that nearest-neighbor pooling on the topic space works better than direct inference on topic models, likely due to the noisy nature of social media text.

Social media popularity and user influence is a very active research area, and our study is a special cases on visual meme. Ginsberg *et al.* [20] showed that web query popularity is correlated with real-world trends such as flu. For text memes, Yang and Leskovec [42] proposes a linear model to predict meme volume, and further quantifies the temporal shape of meme evolution [43]. Other factors that influence popularity include user and the underlying network [44] the nature of the topic [34], and sentiment associated with the message [6]. For views on YouTube, Crane and Sornette [17] characterize the driving mechanisms of YouTube views as driven by external events or internal means (virality). Szabo and Huberman [38] used early video views to predict longer-term view counts. Borghol *et al.* [11] studied whole-clip video clones to quantify content-agnostic factors that influence video views. For describing user roles in information diffusion, Basky *et al.* [5] describes early adopters and influencers for spreading user-created content in virtual communities, Saez-Trumper *et al.* [35] defines trend-setter using graph metrics. Part of our contribution is in demonstrating evidence that fine-grained content features are effective for predicting popularity at individual message level (within a topic).

IX. Conclusion

In this paper, we proposed visual memes for tracking and monitoring of real-world events on content sharing sites like YouTube. We described a system for monitoring event traces in social video, and proposed a scalable algorithm for extracting visual memes with high accuracy. Our system shows that detecting visual memes at a large scale is feasible. Furthermore, we have observed significant remix-behavior in videos discussing news events (up to 70% authors and 60% videos), and observed different user roles in propagating information. We designed a cross-modal matching (CM^2) method for annotating visual memes, and have observed that social network and content features both contribute to better predictions of meme popularity. We are releasing an open-source implementation of the meme-detection algorithm, along with the list of YouTube video ids and the meme detection results on the project page [2].

We would like to point out some limitations of this work. Patterns on two long-running news topics are intended as case studies—there is more work needed before the conclusions generalize to all events or to online video-making in general. The scope and operational definition of video memes in this work is limited to video remixing that largely preserves the visual appearance and semantic content. One can potentially consider other definition of visual memes with larger variations in both appearance and semantics, such as "Dancing Matt",[2] a YouTube phenomena with an iconic person and action performed at different locations. The quantitative evaluation of video meme detection is based on annotations from a single trained annotator; the reliability of the annotation task is unknown.

Future work can take several directions. For example, we may leverage visual memes for better annotations and content search for online videos. We may also examine the sequence and timing of meme shots in relation to popularity and likelihood of a remix. Finally, we are interested in examining visual remixes in video genres other than news.

References

[1] Oxford Dictionaries. [Online]. Available: http://oxforddictionaries.com/definition/english/meme (meme retrieved Jan. 2013).

[2] Project Page and Supplementary Material. [Online]. Available: http://cecs.anu.edu.au/~xlx/proj/visualmemes.html.

[3] YouTube: Statistics. [Online]. Available: http://www.youtube.com/t/press_statistics (retrieved Sep. 2012).

[4] K. Atkinson, Official 12 Dicts Package [Online]. Available: http://wordlist.sourceforge.net (retrieved Mar. 2012).

[5] E. Bakshy, B. Karrer, and L. Adamic, "Social influence and the diffusion of user-created content," in *Proc. ACM EC*, 2009, pp. 325–334.

[6] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an influencer: Quantifying influence on twitter," in *Proc. WSDM*, 2011, pp. 65–74.

[7] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. Blei, and M. Jordan, "Matching words and pictures," *J. Mach. Learn. Res.*, vol. 3, pp. 1107–1135, 2003.

[8] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and K. Ross, "Video interactions in online video social networks," *ACM Trans. Multimedia Comput., Commun., Applicat. (TOMCCAP)*, vol. 5, no. 4, pp. 30:1–25, 2009.

[9] J.-I. Biel and D. Gatica-Perez, "Voices of vlogging," in *Proc. AAAI ICWSM*, 2010, vol. 5.

[10] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–, 2003.

[11] Y. Borghol, S. Ardon, N. Carlsson, D. Eager, and A. Mahanti, "The untold story of the clones: Content-agnostic factors that impact YouTube video popularity," in *Proc. SIGKDD*, 2012, pp. 1186–1194.

[12] U. Brandes, "A faster algorithm for betweenness centrality," *J. Math. Sociol.*, vol. 25, no. 2, pp. 163–177, 2001.

[13] J. Burgess and J. Green, "YouTube: Online video and participatory culture," , 2009.

[14] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi, "Measuring user influence in twitter: The million follower fallacy," in *Proc. ICWSM*, 2010.

[15] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system," in *Proc. ACM IMC*, 2007, pp. 1–14.

[16] C.-C. Chang and C.-J. Lin, LIBSVM: A Library for Support Vector Machines, 2001. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[2] http://en.wikipedia.org/wiki/Matt_Harding

[17] R. Crane and D. Sornette, “Robust dynamic classes revealed by measuring the response function of a social system,” *Proc. Nat. Acad. Sci.*, vol. 105, no. 41, pp. 15649–, 2008.

[18] M. De Choudhury, H. Sundaram, A. John, and D. D. Seligmann, “What makes conversations interesting?: Themes, participants and consequences of conversations in online social media,” in *Proc. WWW*, 2009, pp. 331–340.

[19] B. A. Galler and M. J. Fisher, “An improved equivalence algorithm,” *Commun. ACM*, vol. 7, no. 5, pp. 301–303, 1964.

[20] J. Ginsberg, M. Mohebbi, R. Patel, L. Brammer, M. Smolinski, and L. Brilliant, “Detecting influenza epidemics using search engine query data,” *Nature*, vol. 457, no. 7232, pp. 1012–1014, 2008.

[21] M. Gladwell*, The Tipping Point: How Little Things Can Make a Big Difference Little*. New York, NY, USA: Little, Brown and Co., 2000.

[22] R. Hong, J. Tang, H.-K. Tan, S. Yan, C.-W. Ngo, and T.-S. Chua, “Beyond search: Event driven summarization for web videos,” *ACM Trans. Multimedia Comput., Commun., Applicat. (TOMCCAP)*, 2011.

[23] J. Huang, S. Kumar, M. Mitra, W. Zhu, and R. Zabih, “Spatial color indexing and applications,” *Int. J. Comput. Vision*, vol. 35, no. 3, Dec. 1999.

[24] M. Kendall, “A new measure of rank correlation,” *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938.

[25] J. R. Kender, M. L. Hill, A. P. Natsev, J. R. Smith, and L. Xie, “Video genetics: A case study from YouTube,” in *Proc. ACM Multimedia*, 2010, pp. 1253–1258.

[26] L. Kennedy and S.-F. Chang, “Internet image archaeology: Automatically tracing the manipulation history of photographs on the web,” in *Proc. ACM Multimedia*, 2008.

[27] H. Kwak, C. Lee, H. Park, and S. Moon, “What is twitter, a social network or a news media?,” in *Proc. WWW*, 2010.

[28] J. Leskovec, L. Backstrom, and J. Kleinberg, “Meme-tracking and the dynamics of the news cycle,” in *Proc. KDD*, 2009.

[29] T. Liu, C. Rosenberg, and H. A. Rowley, “Clustering billions of images with large scale nearest neighbor search,” in *Proc. IEEE WACV*, 2007.

[30] D. Machin and A. Jaworski, “Archive video footage in news: Creating a likeness and index of the phenomenal world,” *Visual Commun.*, vol. 5, no. 3, pp. 345–366, 2006.

[31] C. Manning, P. Raghavan, and H. Schutze*, Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008, vol. 1.

[32] M. Muja and D. G. Lowe, “Fast approximate nearest neighbors with automatic algorithm configuration,” in *Proc. Int. Conf. Computer Vision Theory and Applications*, 2009.

[33] A. Natsev, M. Hill, and J. Smith, “Design and evaluation of an effective and efficient video copy detection system,” in *Proc. ICME Workshop*, 2010.

[34] D. M. Romero, B. Meeder, and J. Kleinberg, “Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter,” in *Proc. WWW’11*, 2011, pp. 695–704.

[35] D. Saez-Trumper, G. Comarela, V. Almeida, R. Baeza-Yates, and F. Benevenuto, “Finding trendsetters in information networks,” in *Proc. SIGKDD ACM*, 2012, pp. 1014–1022.

[36] P. Schmitz, P. Shafton, R. Shaw, S. Tripodi, B. Williams, and J. Yang, “International remix: Video editing for the web,” in *Proc. ACM Multimedia*, 2006, pp. 178–.

[37] P. Snickars and P. Vonderau, The YouTube Reader, National Library of Sweden, 2010.

[38] G. Szabo and B. A. Huberman, “Predicting the popularity of online content,” *Commun. ACM*, vol. 53, pp. 80–88, Aug. 2010.

[39] H.-K. Tan, X. Wu, C.-W. Ngo, and W.-L. Zhao, “Accelerating near-duplicate video matching by combining visual similarity and alignment distortion,” in *Proc. ACM Multimedia*, 2008, pp. 861–.

[40] X. Wu, C.-W. Ngo, A. G. Hauptmann, and H.-K. Tan, “Real-time near-duplicate elimination for web video search with content and context,” *IEEE Trans. Multimedia*, vol. 11, no. 2, pp. 196–207, 2009.

[41] L. Xie, A. Natsev, J. R. Kender, M. Hill, and J. R. Smith, “Visual memes in social media: Tracking real-world news in YouTube videos,” in *Proc. ACM Multimedia*, 2011, pp. 53–62.

[42] J. Yang and J. Leskovec, “Modeling information diffusion in implicit networks,” in *Proc. IEEE ICDM*, 2010, pp. 599–608.

[43] J. Yang and J. Leskovec, “Patterns of temporal variation in online media,” in *Proc. WSDM ACM*, 2011, pp. 177–186.

[44] L. Yang, T. Sun, M. Zhang, and Q. Mei, “We know what @you #tag: does the dual role affect hashtag adoption?,” in *Proc. WWW’12*, 2012, pp. 261–270.

Lexing Xie received the B.S. degree from Tsinghua University, Beijing, China, and the M.S. and Ph.D. degrees from Columbia University, respectively. She is Senior Lecturer in Computer Science at the Australian National University, she was with the IBM T.J. Watson Research Center, Hawthorne, NY from 2005 to 2010. Her recent research interests are in multimedia, machine learning and social media analysis. Her research has won six conference paper awards. She plays an active role in editorial and organizational roles in the multimedia community.

Apostol (Paul) Natsev is a software engineer and manager in the video content analysis group at Google Research. Previously, he was a research staff member and manager of the multimedia research group at IBM Research from 2001 to 2011. He received a master’s degree and a Ph.D. in computer science from Duke University, Durham, NC, in 1997 and 2001, respectively. His research interests span the areas of image and video analysis and retrieval, machine perception, large-scale machine learning and recommendation systems. He is an author of more than 70 publications and his research has been recognized with several awards.

Xuming He is a Senior Researcher at National ICT Australia (NICTA) Canberra Lab and an adjunct Research Fellow of Engineering at the Australian National University (ANU). He received the B.S. and M.S. degrees in electrical engineering from Shanghai Jiao Tong University in 1998 and 2001, and Ph.D. degree in computer science from the University of Toronto in 2008. He held a postdoctoral position at the University of California at Los Angeles (UCLA). His research interests include image segmentation and labeling, visual motion analysis, vision-based navigation, and undirected graphical models.

John R. Kender is Professor of Computer Science at Columbia University. His research interests are in the use of statistical and semantic methods for navigating through collections of videos, particularly those showing human activities. He received his Ph.D. from Carnegie Mellon University in 1980, specializing in computer vision and artificial intelligence, and he was the first professor hired in its Robotics Institute. Since 1981, he has been one of the founding faculty of the Department of Computer Science of Columbia University. He has been awarded several teaching awards, and holds multiple patents in computer vision, video analysis, video summarization, and video browsing.

Matthew Hill is a Senior Software Engineer at the IBM T. J. Watson Research Center in Yorktown Heights, NY, where he has worked on systems such as IMARS, the IBM Multimedia Analysis and Retrieval System. Prior to that he was a Software Engineer at Cognex Corporation in Natick, MA, working on industrial machine vision software called PatMax. He holds an M.S. in computer science from Columbia University and a B.A. from Cornell University.

John R. Smith is Senior Manager, Intelligent Information Management Dept, IBM T. J. Watson Research Center. He leads IBMs research in multimedia information retrieval including image/video content extraction, multimedia content-based search, video event detection and retrieval, and social media analysis. He is principal investigator for IBM Multimedia Analysis and Retrieval System (IMARS) and is a long-time participant in the NIST TRECVID video retrieval evaluation. Dr. Smith served as Chair, ISO/IEC MPEG Multimedia Description Schemes Group from 2000–2004 and led the development of multiple parts of the MPEG-7 Multimedia Content Description Standard and MPEG-21 Digital Framework Standard. Dr. Smith is currently Editor-in-Chief of IEEE Multimedia and is a Fellow of IEEE.