

Inferring Social Relations from Visual Concepts

Lei Ding and Alper Yilmaz
Photogrammetric Computer Vision Lab
The Ohio State University, Columbus, OH 43210
leiding326@gmail.com, yilmaz.15@osu.edu

Abstract

In this paper, we study the problem of social relational inference using visual concepts which serve as indicators of actors' social interactions. While social network analysis from videos has started to gain attention in the recent years, the existing work either uses proximity or co-occurrence statistics, or exploit a holistic model of the scene content where the relations are assumed to stay constant throughout the video. This work permits changing relations and argues that there exists a relationship between the visual concepts and the social relations among actors, which is a fundamentally new concept in computer vision. Specifically, we leverage the existing large-scale concept detectors to generate concept score vectors to represent the video content, and we further map them to grouping cues that are used to detect the social structure. In our framework, a probabilistic graphical model with temporal smoothing provides a means to analyze social relations among actors and detect communities. Experiments on Youtube videos and theatrical movies validate the proposed framework.

1. Introduction

Inferring social relations among actors in a video refers to the process of associating the information content of the video to interactions among the actors in it. These social relations are typically encoded in a weighted adjacency matrix depicting the underlying social network which takes actors as its vertices. A traditional goal in social network analysis is to detect communities of actors, such that actors from the same community share strong bonds [20]. In the recent years, we have seen a handful of studies in the literature which exploit this theme [25, 21, 5]. These studies, however, focus only on computing the social relations using proximity heuristics [25], co-occurrence statistics [21], or holistic scene content [5]. In this paper, we take the first step to infer social relations with the existing body of work on visual concept detection. Following the discussion in [18], we believe social relations, when used with other observa-

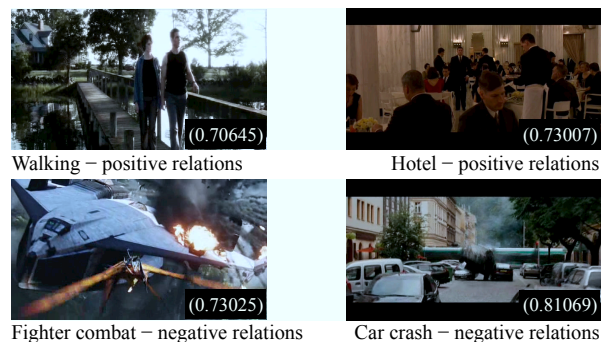


Figure 1. Visual concepts are indicative of social relations among actors. The first row shows concepts that are generally related to actors from the same social community, while the second row shows the opposite. In parentheses, we report the concept detection scores which are bounded between $[0, 1]$.

tions, provide significant contextual information that will aid disambiguating hard computer vision problems, such as object recognition and action recognition.

Our main conjecture is that visual concept detection, compared to low-level video information, provides useful semantic features for inferring social relations. For example, individuals involved in fighting on a battle scene tend to be enemies, while individuals jogging leisurely together tend to be friends. In Figure 1, we show a set of examples depicting this observation. In this figure, different visual concepts are linked to social relations among actors in either the same or different (sometimes rival) communities. The relations between locations, actions and social interactions have recently been studied in other research areas [3, 14]; however, to the best of our knowledge, such intuition has not been considered in the computer vision field.

In the recent years, semantic visual concept detection, such as snow, river, car racing, etc., has become a popular tool in indexing and retrieval of visual content in large image and video collections [22, 26, 19, 4, 6]. Recently, researchers, information analysts, and ontology specialists

jointly defined the LSCOM ontology [12], which includes more than 834 *useful, feasible and observable* visual concepts that are related to events, objects, locations, people, and programs which are found in broadcast news videos. From among these semantic visual concepts, Yanagawa *et al.* [23] have provided support vector machine (SVM) detectors for 374 trained concepts, which are referred to as Columbia-374. Due its public availability and broadness, we have adopted the Columbia-374 as our base detectors.

Given the visual concepts detected using Columbia-374, we map the video content to a *grouping cue* at each video scene, and learn social network representations using actor occurrence patterns together with the inferred grouping cues. Once social networks are estimated, we use a probabilistic graphical model as the analytical tool, which poses community detection as a probabilistic inference problem. In comparison to state-of-the-art, the proposed approach allows changes in social relations over time and assumes temporal smoothness in such relations. This treatment leads to better accuracy in inferring the relations and enables the new functionality of analyzing changing relations.

The two main contributions made in this paper include inference of social relations from visual concepts, and introduction of a probabilistic model with temporal smoothing for discovering social communities. The remainder of this paper is organized as follows. We will first give a brief overview of related research. In Section 2, we will discuss the use of visual concepts, followed by Section 3 on learning and analyzing social networks. We report experimental results in Section 4, and conclude in Section 5.

Related Work Social networks have long been a topic of research in sociology, which have been predominantly used to detect communities and understand the flow of information among their members [20]. Recently, other research fields, which include data mining, computer vision and multimedia analysis, have started using social networking to solve problems in their respective domains [24, 25, 21, 7]. For instance, Eagle and Pentland [7] have exploited the mobile phone usage log-entries (non-visual) to infer social networks. As videos become ubiquitously available, we believe the research proposed in this paper will open new directions for studying social phenomena from videos.

Extracting social content from video is a challenging task, primarily, due to the gap between low-level features and high-level social interactions. In context of surveillance videos, Yu *et al.* [25] define social relations between actors using a proximity heuristic. While not necessarily representing social phenomena, resulting relations provide their approach a basis to detect communities using a traditional social networking tool referred to as the modularity cut [13]. In a similar fashion, Ge *et al.* [9] define existence of social relations based on proximity and relative velocity between

the objects, which are later used to detect communities by using well-known clustering techniques. For learning communities and leaders in videos, Ding and Yilmaz [5] construct a social network from low-level audiovisual features which relate to learned interactions among actors. The authors use a modified max-min modularity algorithm to extract communities from the social network.

A common trend in social network analysis, including the ones cited above, is to assume that the social networks remain constant and the relations between actors do not change in time. This assumption requires an off-line analysis of the video and does not respect the evolutionary nature of human interactions. We address these limitations by using a probabilistic method which models evolving communities as a stochastic process of actor-community assignments. This treatment of social relations relates to the model proposed in [24]. Our approach, however, considers real-valued links that result from general modes of interactions in a social setting. The probabilistic characteristic provides our approach with the ability to elegantly generate more than two communities compared to the modularity cut, which requires a heuristic iterative process [13].

2. Using the Visual Concepts

Our work is based on the explicit use of visual concepts for inferring social structures among the actors in a video. In this section, we will first discuss the large repository of base detectors that we adopt to represent the video content. Following this discussion, we will introduce the approach taken to map concept detection scores to grouping cues. The mapping is performed at the scene-level and provides a means to characterize the relations among actors within that video scene.

2.1. Base Concept Detectors

For detecting base concepts, we use the three low-level features exploited in [23], which include the color, texture and edges computed from keyframes in a video. Specifically, for the *grid color moment* (GCM) feature, we extract the first 3 moments of the 3 channels in the CIE LUV color space over 5×5 fixed grid partitions, and aggregate the features into a single 225-dimensional feature vector. The texture is modeled by the *Gabor texture* (GT) feature, which we extract by taking 4 scales and 6 orientations of Gabor transformations and use their means and standard deviations. This process provides a texture feature in the form of a vector with 48 dimensions. The edge content in the image is modeled using the *edge direction histogram* (EDH), which is composed of 73 dimensions corresponding to 72 bins of edge direction quantized at 5 degree intervals and 1 bin for non-edge points.

A video is composed of multiple scenes which may contain individuals performing activities for a common cause.

In our approach, we use the temporal extent of a scene to extract observations relating to the social content and consider that the visual content in a scene is represented by its keyframes. Following the extraction of D keyframes using [15], we compute low-level features for each keyframe i . These features are then used to estimate a normalized score from the SVMs which are independently trained on each low-level feature¹: $f_{i,j}$ for three feature types $j = 1, 2, 3$. The average of the three scores $\bar{f}_i = \frac{1}{3}(f_{i,1} + f_{i,2} + f_{i,3})$ is used as the overall score for the i^{th} keyframe. Finally, we compute the visual concept score by max-pooling over all keyframes within the scene bounds, $\max_i \bar{f}_i$. This process, when performed for all 374 visual concepts, results in a 374-dimensional semantic vector representing the scene. Each element of the semantic vector provides the confidence score corresponding to a semantic concept. This mid-level representation has previously been shown to abstract the visual content in videos [22] and is used as the basis for inferring social relations in this paper.

2.2. Mapping Semantic Concepts to Grouping Cues

Let us assume that a video \mathbb{V} is composed of M scenes, $s_1, s_2 \dots s_M$, each of which contains a set of actors \mathcal{C}_i and has an associated grouping cue $\beta_i \in [-1, +1]$. The grouping cue serves as a basis to decide whether the actors co-occurring in the scene belong to the same ($\beta_i > 0$) or different ($\beta_i < 0$) communities. In our setting, the larger the absolute value of β_i is, the more stringent the corresponding constraints are. Note that the concept of grouping cue generalizes the adverseness measure as used in [5], as we do not stipulate the types of videos or social relations.

Clearly, not all the dimensions of the semantic vectors are equally informative towards social relations. Additionally, detecting some of the visual concepts may be unsatisfactory due to their large variability or relatively small spatiotemporal extents. To address this issue, we use a supervised dimension reduction method known as kernel local Fisher discriminant analysis (KLFDA) [17], which effectively combines the ideas of Fisher discriminant analysis and locality preserving projection [8, 10]. It has an analytic form of the embedding transformation and the solution can be computed by solving a generalized eigenvalue problem. Due to the page limit, we refer the reader to the details in [17]. In Figure 2, we show several examples of this algorithm on synthetic data to demonstrate its capabilities in reducing dimensions based on the data structure and given labels. By applying this data-dependent transform, we derive a more informative and compact representation, which is a d -dimensional vector for each video scene, for learning the social relations.

In order to estimate the grouping cues β_i from the d -

¹A normalized score is the raw decision value from an SVM transformed by a logistic function.

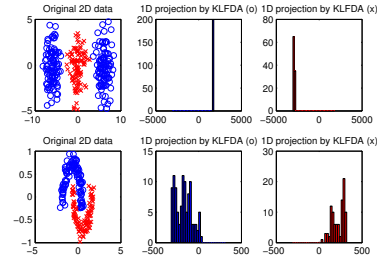


Figure 2. Synthetic examples show the power of supervised dimension reduction. Displayed are the histograms of the projected 1D values for the two classes. After the projection, only useful information towards class labeling is kept.

dimensional transformed semantic vectors, we use support vector regression (SVR). The goal is to find a function $g(\cdot)$ that has at most ϵ deviation from the labeled targets for all the training data, and at the same time is as flat as possible. It is shown in [16] that the final decision function can be written as: $\beta_i = g(s_i) = \sum_{j=1}^L (\alpha_j - \alpha_j^*) \mathcal{K}_{l_j, i} + b$, where α_i and α_i^* are the Lagrange multipliers for labeling constraints, b is an offset, L is the number of labeled scene examples, and l_j is the index for the j^{th} labeled scene example. In our problem domain, the kernel \mathcal{K} is chosen as a radial basis function kernel over the concept score vectors. This kernel together with training scenes and their grouping cues $\beta_i = +1$ (scene with members of same community) and $\beta_i = -1$ (scene with members from different communities) leads to grouping constraints for a novel video. This is achieved by estimating the corresponding β_i using the regression learned from labeled video scene examples from other videos in the training set.

3. Social Relational Inference

The proposed model for social relational inference from videos relies on estimated social networks from which social communities are detected. Let each video contain a number of partitions, each of which is composed of a number of scenes. We build social networks with connectivity matrix W^t to represent social relations from the start of the video until partition t . This treatment honors the evolving nature of relational formation and allows social networks that can change based on past observations.

Specifically, social networks are learned by means of a Gaussian process based affinity learning method. The actor-scene appearance matrix $A = \{a_{ij}\}$, where $a_{ij} = 1$ if actor c_j appears in scene s_i , and estimated β_i 's up to the end of partition t of the video (or equivalently $\phi(t)^{\text{th}}$ scene) are used to infer the membership vector \mathbf{f} of the actors. It can be verified that $P(\mathbf{f}|A, \beta) \propto \exp(-\frac{1}{2}\mathbf{f}^T W^{t-1} \mathbf{f})$ is a Gaussian process with zero mean. We let $W^t = \{w_{ij}^t\}$, and use $w_{ij}^t = E\{f_i f_j | A, \beta\}$ as the learned affinity between

Algorithm 1 Community detection from social networks

Input: Social affinity matrices W_T .

Output: $Z_T^* \approx \arg \max P(Z_T|W_T)$ with the post-convergence Z_T samples generated below.

Randomly initialize Z_T .

for $r = 1$ to M_s **do**

for each time step t **do**

for each actor i **do**

 Compute the distribution $P(z_i^t|Z_{T \setminus \{i,t\}}, W_T)$ using the current Z_T .

 Sample the community assignment z_i^t from the distribution above, and update the corresponding entries of Z_T .

end for

end for

end for

the actors c_i and c_j . In this formulation, it follows that $W^t = (M^t)^{-1}$, where the elements of $M^t = \{m_{ij}^t\}$ are:

$$m_{ij}^t = \sum_{k \leq \phi(t): c_i, c_j \in s_k} -\text{sgn}(\beta_k) \alpha_k^2 \beta_k^2, \quad (1)$$

for $i \neq j$ where $\text{sgn}(\cdot)$ is a sign function; otherwise,

$$m_{ii}^t = 1 + \sum_{l \neq i} \sum_{k \leq \phi(t): c_i, c_l \in s_k} \alpha_k^2 \beta_k^2. \quad (2)$$

In these equations, $\alpha_k = \exp(-\rho(\phi(t) - k))$, where ρ is a decay parameter emphasizing more recent video content.

We proceed by separating the learned affinity matrix W^t into *principal* $U^t = \{u_{ij}^t\}$ and *complementary* $V^t = \{v_{ij}^t\}$ affinity matrices where $W^t = U^t - V^t$, $u_{i,j}^t = w_{i,j}^t$ for positive entries of W^t , and $v_{i,j}^t = -w_{i,j}^t$ for negative entries of W^t . U^t and V^t respectively represent the *relatedness* and *unrelatedness* between actors in terms of community memberships. Additionally, in order to suit the probability mass functions presented next, values of U^t and V^t are quantized into n_b uniform bins, starting from 0. Thus, their values are converted to integers within the range $[0, n_b - 1]$.

Community detection deals with associating each one of the n actors to K different communities at each time step t . As illustrated in Figure 3, at time instant t , the hidden quantity to estimate is the binary community assignment matrix $Z^t = \{z_{ik}^t\}$, which is an $n \times K$ matrix. In this setting, z_i^t denotes the membership of actor i at time t , such that, $z_{ik}^t = 1$ for $k = z_i^t$, and $z_{ik}^t = 0$ for $k \neq z_i^t$. In the figure, the observation at each time step is denoted by W^t which is estimated ahead of time from co-occurrence patterns and learned visual concepts. Intuitively, the model works by maximizing the agreement between Z^t and W^t . It also encourages the nearby Z^t estimates to evolve smoothly in a temporal order. Specifically, the probabilistic relations among variables are as follows,

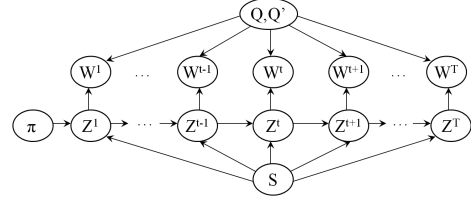


Figure 3. The model for detecting social relations, which outputs detected community memberships of actors at each time step.

- $z_i^1|\pi$ is a multinomial distribution, $P(z_i^1|\pi) = \prod_k \pi_k^{z_{ik}^1}$.
 - $z_i^t|z_i^{t-1}, S$ follows a multinomial distribution, $P(z_i^t|z_i^{t-1}, S) = \prod_{k,l} S_{kl}^{z_{ik}^{t-1} z_{il}^t}$, where the transition probability matrix $S \in \mathbb{R}^{K \times K}$. That is, if actor i belongs to community k at $t-1$, then at time t , he will remain in the same community with probability S_{kk} , or he will change to another community l with probability S_{kl} .
 - $u_{i,j}^t|z_i^t, z_j^t, Q$ follows a geometric distribution, $P(u_{i,j}^t|z_i^t, z_j^t, Q) = \prod_{k,l} (Q_{kl}^{u_{ij}^t} (1 - Q_{kl}))^{z_{ik}^t z_{jl}^t}$.
 - $v_{i,j}^t|z_i^t, z_j^t, Q'$ follows a geometric distribution, $P(v_{i,j}^t|z_i^t, z_j^t, Q') = \prod_{k,l} (Q'_{kl}^{v_{ij}^t} (1 - Q'_{kl}))^{z_{ik}^t z_{jl}^t}$.
- The matrices $Q, Q' \in \mathbb{R}^{K \times K}$ are link probability matrices respectively for principal and complementary affinities, where diagonal and off-diagonal elements respectively refer to within-community link probabilities and between-community link probabilities.

One possibility to estimate the model parameters π, S, Q and Q' is to use training data. Point estimation for these parameters, however, is less stable and reliable, especially in the case when data samples are insufficient and noisy [11]. Following this observation, we use Bayesian inference and attach conjugate prior distributions to the unknown parameters [24]. In this way, Dirichlet distributions are assumed for π and S , and Beta distributions are assumed for Q and Q' as follows, where $\Gamma(\cdot)$ is a Gamma function:

$$P(\pi) = \frac{\Gamma(\sum_k \gamma_k)}{\prod_k \Gamma(\gamma_k)} \prod_k \pi_k^{\gamma_k - 1}, \quad (3)$$

$$P(S) = \prod_k \frac{\Gamma(\sum_l \mu_{kl})}{\prod_l \Gamma(\mu_{kl})} \prod_l S_{kl}^{\mu_{kl} - 1}, \quad (4)$$

$$P(Q) = \prod_{k,l \geq k} \frac{\Gamma(\alpha_{kl} + \beta_{kl})}{\Gamma(\alpha_{kl})\Gamma(\beta_{kl})} Q_{kl}^{\alpha_{kl} - 1} (1 - Q_{kl})^{\beta_{kl} - 1}, \quad (5)$$

$$P(Q') = \prod_{k,l \geq k} \frac{\Gamma(\alpha'_{kl} + \beta'_{kl})}{\Gamma(\alpha'_{kl})\Gamma(\beta'_{kl})} Q'_{kl}{}^{\alpha'_{kl} - 1} (1 - Q'_{kl})^{\beta'_{kl} - 1}. \quad (6)$$

Let $W_T = \{W^i\}_{i=1}^T$, $Z_T = \{Z^i\}_{i=1}^T$ and $\theta = \{\pi, S, Q, Q'\}$. In a video some actors may appear at a later time compared to other actors. In order to handle such situations, we denote the first time step that actor i appears as

τ_i . We also define the following auxiliary function:

$$\begin{aligned}
L(\alpha, \beta, w) = & \Pi_{k,l>k} B(\alpha_{kl} + \sum_{t=1}^T \sum_{i:\tau_i \leq t} \sum_{j \neq i: \tau_j \leq t} (z_{ik}^t z_{jl}^t + z_{il}^t z_{jk}^t), \\
& \beta_{kl} + \sum_{t=1}^T \sum_{i:\tau_i \leq t} \sum_{j \neq i: \tau_j \leq t} w_{ij}^t (z_{ik}^t z_{jl}^t + z_{il}^t z_{jk}^t)) \\
& \times \Pi_k B(\alpha_{kk} + \frac{1}{2} \sum_{t=1}^T \sum_{i:\tau_i \leq t} \sum_{j \neq i: \tau_j \leq t} (z_{ik}^t z_{jk}^t + z_{ik}^t z_{jk}^t), \\
& \beta_{kk} + \frac{1}{2} \sum_{t=1}^T \sum_{i:\tau_i \leq t} \sum_{j \neq i: \tau_j \leq t} w_{ij}^t (z_{ik}^t z_{jk}^t + z_{ik}^t z_{jk}^t)), \quad (7)
\end{aligned}$$

where $B(\cdot)$ is a Beta function. Given the above distributions, the overall likelihood is computed by:

$$\begin{aligned}
P(W_T, Z_T) = & \int \left(\prod_{t=1}^T P(W^t | Z^t, Q, Q') \right. \\
& \left. \prod_{t=2}^T P(Z_{i:\tau_i \leq t-1}^t | Z_{i:\tau_i \leq t-1}^{t-1}, S) \prod_{i=1}^n P(z_i^{\tau_i} | \pi) P(\theta) \right) d\theta \\
\propto & \Pi_k \Gamma \left(\sum_{i=1}^n z_{ik}^{\tau_i} + \gamma_k \right) L(\alpha, \beta, u) L(\alpha', \beta', v) \\
& \times \Pi_k \frac{\prod_i \Gamma(\sum_{t=2}^T \sum_{i:\tau_i \leq t-1} z_{ik}^{t-1} z_{il}^t + \mu_{kl})}{\Gamma(\sum_{t=2}^T \sum_{i:\tau_i \leq t-1} z_{ik}^{t-1} + \sum_l \mu_{kl})}. \quad (8)
\end{aligned}$$

In the case when all actors appear at $t = 1$, we have $\tau_i = 1$ for all i , such that the above equation can be simplified. For details on Gibbs sampling used for our social relational inference algorithm we refer the reader to Algorithm 1. In short, the key is to sample from $P(z_i^t | Z_{T \setminus \{i,t\}}, W_T)$, where $Z_{T \setminus \{i,t\}}$ denotes all the community memberships except actor i 's at time step t . This quantity is proportional to $P(Z_T, W_T)$ given fixed values of all other community assignments. After the inference algorithm stops, the proposed approach provides the complete set of Z_T estimates. In other words, the corresponding communities become detected. The time complexity of the whole process is $O(M_s T^2 n^3)$, where M_s is the number of Gibbs iterations.

4. Experiments

In order to demonstrate the performance of the proposed framework, we experiment with two types of videos: theatrical movies and Youtube videos. In particular, the experiments on theatrical movies provide us with the ability to compare against [5]. Experiments on Youtube videos are exercised to test the approach with data that introduce new challenges, such as low quality and short duration. Additionally, we have expanded the movie dataset provided in [5] to 20 movies².

²The additional ten movies cover a variety of genres and include: (1) Austin Powers - The International Man of Mystery (1997); (2) The Dark Knight (2008); (3) Harry Potter and the Goblet of Fire (2005); (4) The Italian Job (2003); (5) Minority Report; (6) Office Space (1999); (7) X-Men: The Last Stand (2006); (8) Avatar (2009); (9) The Bourne Ultimatum (2007); (10) Pirates of the Caribbean: Curse of the Black Pearl (2003). The datasets are available at <http://dpl.ceegs.ohio-state.edu/resources.php>.

Most informative	Rates	Least informative	Rates
NaturalDisasters	68.5%	Cloverleaf	50.0%
Shooting	68.2%	SingleFamilyHomes	54.1%
Groom	68.1%	Girl	55.1%
Moonlight	68.1%	WaterTower	55.8%
Cityscape	68.1%	Non-usNationalFlags	55.8%
Ship	68.0%	Farms	56.5%
Beach	67.9%	Computers	56.8%
BoatShip	67.8%	Camera	57.0%
Cheering	67.8%	Bicycle	57.0%
Sitting	67.8%	CigarBoats	57.0%

Table 1. Average accuracy using a single visual concept for grouping cue estimation using the movie dataset. **Left:** 10 most informative concepts with the highest accuracy. **Right:** 10 least informative concepts with the lowest accuracy.

Preliminary Analysis Before performing social inference, we first validate the information retained by visual concept detection for social relational analysis. Since each visual concept detector provides different information, we inspect the most and the least informative concepts by performing empirical analysis on the 20-movie dataset using the following procedure.

For each of the 374 concept detectors, we perform kernel density estimation of the concept scores for scenes where grouping cue labels equal +1. Similarly, we estimate densities for scenes with -1 grouping cues. For each visual concept, this procedure provides two one-dimensional density functions. According to the maximum likelihood (ML) rule, we compute the classification accuracy by applying the two estimated density functions on the two classes of concept scores respectively. Therefore, each visual concept i has a corresponding accuracy rate p_i , averaged over the two classes. The larger this value is, the more information it contains for social relational analysis. We tabulate the outcome of this analysis in Table 1 for the 10 most (left) and 10 least (right) informative concepts. As can be observed, concepts like “beach” and “shooting” are informative, since they contain information relating to social phenomena. On the other hand, concepts like “camera” and “bicycle” do not contain as much information, due to the fact they are not easy to recognize and do not relate as much to social interactions. “Farms” is a concept uncovered in our movie dataset, and thus it is less informative. While a semantic concept vector carries more information compared to the individual concepts, our analysis demonstrates the usefulness of the concepts for relational inference.

Movie dataset We perform two sets of experiments on movies. The first experiment uses the 10-movie dataset provided in [5] in order to compare with the results therein. The second experiment is performed on the larger dataset with 20 movies to perform comparative analysis of a set of



Figure 4. Detected communities for 20 movies. Actors are arranged in communities (C_1 , C_2 and C_3 from top to bottom) according to the movie plots. Actors incorrectly associated with communities are displayed in colored shades. For those actors, computed community indices are marked next to their names when the movie contains more than two communities.

Methods	Prec(+)	Prec(-)	Prec(ave)	F_1
Baseline [5]	—	—	78.2%	76.0%
Proposed-10 (d=50)	83.1%	85.0%	84.1%	81.9%
Proposed-20 (d=100)	76.9%	85.7%	81.3%	80.7%
Proposed-20 (d=50)	77.8%	88.0%	82.9%	81.1%
Modularity-20	—	—	—	77.5%
Spectral clustering-20	—	—	—	78.5%
PCA-20 (d=100)	77.1%	80.7%	78.9%	75.6 %
PCA-20 (d=50)	76.7%	79.5%	78.1%	75.3 %
Base concept-20	75.5%	78.5%	77.0%	74.9%
No concept-20	43.6%	35.0%	39.3%	41.0%

Table 2. Comparative analysis against other approaches. The measures used are as follows: Prec(+): precision of β estimates for the positive class; Prec(-): precision of β estimates for the negative class; Prec(ave): average precision; and F_1 measure for community detection averaged over all videos, which is the final score for social relational inference.

different methods for grouping cue labeling and community detection. The parameter used for videos are as follows: $\gamma_k = 10$, $\mu_{kk} = 10$, $\mu_{kl} = 1 (k \neq l)$, $\alpha_{kk} = 100$, $\alpha_{kl} = 1 (k \neq l)$, $\beta_{kl} = 10$, $\alpha'_{kk} = 1$, $\alpha'_{kl} = 100 (k \neq l)$, $\beta'_{kl} = 10$, $\rho = 0.01$, $M_s = 500$, $D = 10$, $\phi(t) = 20t$. The movie scenes and actor-scene appearance matrices are detected using the method described in [5]. Quantitative analysis tabulated in Table 2 is performed on the following approaches, where the number 10 or 20 respectively refers to the smaller and larger dataset, and d refers to transformed dimensionality:

- Baseline approach [5]: SVR and modularity cut are used for learning and analyzing social networks;

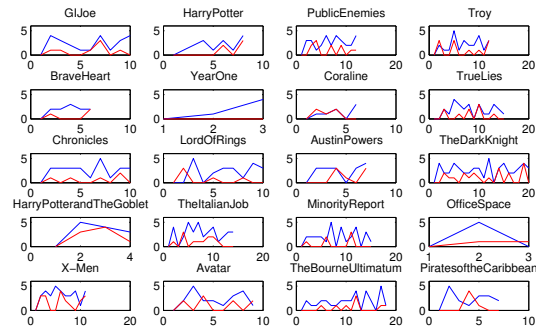


Figure 5. The number of actors changing community membership as a function of time. Red curves describe the results from our framework. Blue curves show the results from a variant of our framework where the temporal links are removed between all pairs of Z^t and Z^{t+1} in the graphical model.

- Proposed: the framework detailed in this paper. The semantic vectors are transformed using KLFDA;
- Modularity variant: modified max-min modularity cut in [5] is used for detecting communities instead of the proposed method;
- Spectral clustering variant: the spectral clustering method in [2] with temporal smoothing is used for detecting communities instead of the proposed method;
- PCA variant: supervised feature extraction (KLFDA) is replaced with principle component analysis;
- Base concept variant: supervised feature extraction is skipped in the proposed framework;
- No concept variant: the visual concept detection step is removed from our framework. That is, the low-level features are directly mapped to grouping cues.

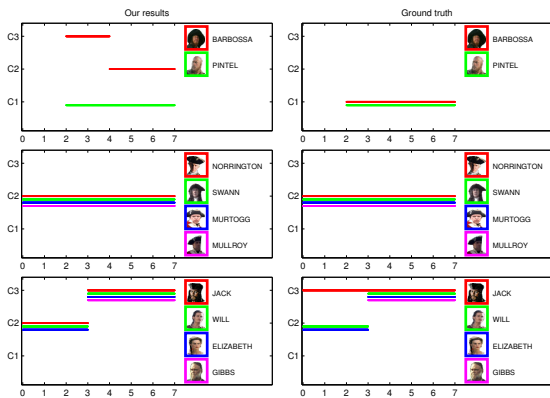


Figure 6. Evolution of communities with respect to time. Community membership is plotted as a function of time. **Left:** automatically computed results. **Right:** ground truth labeling according to movie plots. Best viewed in color.

From the table, it can be seen that the proposed method works better than the baseline approach on both reported measures. On the same 10-movie dataset, our method works considerably better by close to 6% in the final F_1 measure of community detection, with no use of tailored visual features. On the larger set, our proposed method outperforms other alternatives for grouping cue estimation or community detection by appreciable margins. Generated communities from our proposed approach are shown in Figure 4. The actors’ names are arranged in boxes according to ground truth social communities. Actors with incorrectly computed community memberships judging from ground truth are displayed in colored shades. The computed group indices are marked next to the actors’ names when the movie contains more than two communities.

Finally, we show the evolving actor-community associations in Figure 5 for all 20 movies in the dataset. The figure contains plots of the number of actors changing their community membership at each time step. Red curves describe the results from our framework, while blue curves show the results from a variant of our framework where the temporal links are removed between all adjacent pairs of Z^t vertices, and π is connected to all Z^t in the graphical model. It can be seen that our proposed approach with temporal links helps to smooth the social relation estimates, as the number of actors updating their membership tends to be smaller. For detailed analysis on dramatic relational changes, we illustrate an example in Figure 6 for the three communities occurring in the movie titled *Pirates of the Caribbean: Curse of the Black Pearl* (2003). In the figure, the communities are represented as boxes, where the left column shows the computed results and the right column shows the ground truth. Our approach successfully detects the community evolution of actors from C_2 to C_3 .

Youtube dataset In order to demonstrate the generality of our proposed framework, we collected 20 Youtube videos from two categories (soccer game and demonstration, see Figure 7), each containing 10 videos. In this experiment, due to the lack of detailed actor identification, and the infeasibility to compute social communities, we test the performance of grouping cue estimation using visual concepts. The soccer videos contain two competing communities and the demonstration videos contain policemen and demonstrators confronting the policemen.

We consider visual event categories and label them as an intermediate step. For soccer videos, the labeled events are “chasing”, “confronting”, “hugging” and “others”. For demonstration videos, the labeled events are “marching”, “confronting”, “public speaking” and “others”. We assume the video composed of 4 second scenes that respect the shot boundaries. In order to evaluate the estimation of grouping cues at the scene level, we manually generate the ground truth for each video, which labels +1 if only members from one community appear and -1 if members from both communities co-appear³. Learning is carried out using the leave-one-video-out strategy.

In addition to the features introduced earlier (GCM, GT, EDH), we include a motion representation similar to the histogram of oriented optical flows presented in [1]. Around the keyframe of each scene, optical flow orientation histograms are generated in a window of 20 frames. The 8-bin histogram is weighted by the magnitude of observed motion, and provides an 8-dimensional motion feature vector for the video scene. A joint product kernel fusing appearance and motion features provides the similarity between the video scenes. Finally, support vector learning is used to map to visual concepts and then to grouping cues as discussed for the movie experiment.

In Figure 8, we summarize the performance of estimating grouping cues for each type of video in terms of the precision rate, which is the percentage of accurately labeled scenes in each predicted sub-category. In the figure, $\text{Prec}(+)$ refers to precision of the positive class and $\text{Prec}(-)$ refers to that of the negative class. It can be seen that the variant with additional motion features outperforms the variant without motion features, possibly due to rich motion content in these video types. We should note that the relatively low performance of the negative group for demonstration videos is due to the non-discriminative visual features in demonstrations that contain both communities. Collectively, learning grouping cues from event modeling works well for both video types with average precision rates at 84.8% and 72.0% respectively for soccer and demonstration videos using the combination of features.

³The ground truth labeling is performed by individuals. The final labels are chosen only when more than two individuals agree.



Figure 7. Interactions within (left two images) and across (right two images) communities have distinctive patterns. The scenes are taken from soccer-game and demonstration videos downloaded from Youtube.

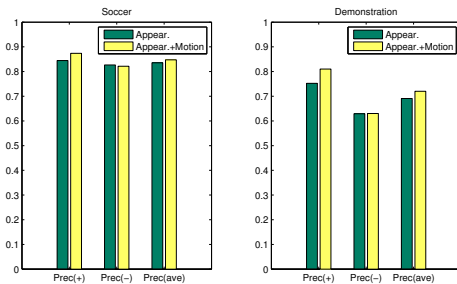


Figure 8. Precision rates of grouping cue labeling for the two types of videos. Listed are Prec(+) (precision of the positive class), Prec(-) (precision of the negative class) and their average.

5. Conclusions

In this paper, we have shown that that social relational inference can benefit from detecting the visual concepts from videos. We have also introduced a temporally smoothed probabilistic model for performing community detection from learned social networks with real-valued links. Experimental results show that the proposed framework has worked well and advanced the state-of-the-art of social relational inference from videos.

References

- [1] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *CVPR*, 2009. 7
- [2] Y. Chi, X. Song, D. Zhou, K. Hino, and B. L. Tseng. Evolutionary spectral clustering by incorporating temporal smoothness. In *KDD*, 2007. 6
- [3] S. Consolvo, I. E. Smith, T. Matthews, A. LaMarca, J. Tabert, and P. Powledge. Location disclosure to social relations: Why, when & what people want to share. In *CHI*, 2005. 1
- [4] L. Ding, Q. Fan, J. Hsiao, and S. Pankanti. Graph based event detection from realistic videos using weak feature correspondence. In *ICASSP*, 2010. 1
- [5] L. Ding and A. Yilmaz. Learning relations among movie characters: A social network perspective. In *ECCV*, 2010. 1, 2, 3, 5, 6
- [6] L. Duan, D. Xu, I. W. Tsang, and J. Luo. Visual event recognition in videos by learning from web data. In *CVPR*, 2010. 1
- [7] N. Eagle, A. Pentland, and D. Lazer. Inferring social network structure using mobile phone data. *PNAS*, 106(36):15274–15278, 2009. 2
- [8] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936. 3
- [9] W. Ge, R. Collins, and B. Ruback. Automatically detecting the small group structure of a crowd. In *WACV*, 2009. 2
- [10] X. He and P. Niyogi. Locality preserving projections. In *NIPS*, 2003. 3
- [11] S. M. Lynch. *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. Springer, New York, 2007. 4
- [12] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE Multimedia*, 13(3), 2006. 2
- [13] M. E. J. Newman. Modularity and community structure in networks. *PNAS*, 103(23):8577–8582, 2006. 2
- [14] A. Pentland. *Honest Signals: How They Shape Our World* (Bradford Books). The MIT Press, 2008. 1
- [15] Z. Rasheed and M. Shah. Scene detection in hollywood movies and tv shows. In *CVPR*, 2003. 3
- [16] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, 2004. 3
- [17] M. Sugiyama. Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis. *Journal of Machine Learning Research*, 8:1027–1061, 2007. 3
- [18] G. Wang, A. Gallagher, J. Luo, and D. Forsyth. Seeing people in social context: Recognizing people and social relationships. In *ECCV*, 2010. 1
- [19] P. Wang, G. D. Abowd, and J. M. Rehg. Quasi-periodic event analysis for social game retrieval. In *ICCV*, 2009. 1
- [20] S. Wasserman, K. Faust, and D. Iacobucci. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994. 1, 2
- [21] C.-Y. Weng, W.-T. Chu, and J.-L. Wu. Rolenet: Movie analysis from the perspective of social networks. *IEEE Transactions on Multimedia*, 11(2):256–271, 2009. 1, 2
- [22] D. Xu and S.-F. Chang. Visual event recognition in news video using kernel methods with multi-level temporal alignment. In *CVPR*, 2007. 1, 3
- [23] A. Yanagawa, S.-F. Chang, L. Kennedy, and W. Hsu. Columbia university’s baseline detectors for 374 Iscom semantic visual concepts. Technical report, Columbia U. 2
- [24] T. Yang, Y. Chi, S. Zhu, Y. Gong, and R. Jin. A bayesian approach toward finding communities and their evolutions in dynamic social networks. In *SDM*, 2009. 2, 4
- [25] T. Yu, S.-N. Lim, K. Patwardhan, and N. Krahnstoeber. Monitoring, recognizing and discovering social networks. In *CVPR*, 2009. 1, 2
- [26] X. Zhou, X. Zhuang, S. Yan, S.-F. Chang, M. Hasegawa-Johnson, and T. S. Huang. Sift-bag kernel for video event analysis. In *MM*, 2008. 1