

Chapter 2

Learning Social Relations from Videos: Features, Models, and Analytics

Lei Ding and Alper Yilmaz

2.1 Introduction

Despite the progress made during the last decade in video understanding, extracting high-level semantics in the form of relations among the actors in a video is still an under-explored area. This chapter discusses a streamlined methodology to learn interactions between actors, construct social networks, identify communities, and find the leader of each community in a video sequence from a sociological perspective. Specifically, we review one of the first studies reported in [8] toward learning such relations from videos using visual and auditory cues. The main contribution can be stated as the association of low-level video features to social relations by means of machine learning mechanisms, including support vector regression and Gaussian processes. The resulting social network is then analyzed to find communities of actors and identifying the leader of each community, which are two of the most important tasks in social network analytics. Furthermore, as an extension to the basic framework, we discuss the relationship between visual concepts and social relations that has been explored in [9]. In this setting, visual concepts serve as mid-level visual representation in inferring social relations and are compared with features employed in the basic framework.

Recently, researchers have devoted countless efforts on understanding the scene content from video by analyzing the object trajectories and finding common motion patterns [2, 7, 13, 22, 23, 44]. Most of these efforts, however, did not go beyond analyzing or grouping trajectories, or understanding individual actions performed by tracked objects [3, 16, 20, 36, 43]. The computer vision community, generally

L. Ding (✉)
The Ohio State University, Boston, MA 02110, USA
e-mail: leiding326@gmail.com

A. Yilmaz
The Ohio State University, Columbus, OH 43210, USA
e-mail: yilmaz.15@osu.edu

speaking, did not consider analyzing the video content from a sociological perspective, which would provide systematic understanding of the roles and activities performed by actors based on their relations. In relation to the existing body of work on action or event recognition and analysis, better analyzed social relations, when used with other feature observations, can provide useful contextual information to aid in disambiguating hard-to-recognize actions, events, or objects [29, 38].

In sociology, social structures are believed to be best represented and analyzed using a social network [39]. Social network analytics views social relations in terms of a network consisting of vertices and edges. The vertices represent individual actors within the network, and the edges denote the relations between the actors. The resulting graph structure provides a means to detect and analyze communities in the network. The communities are traditionally detected based on the connectivity between the actors using social network tools, such as the popular modularity algorithm [28]. Social network analytics has recently attracted much interest in the fields of data mining [30, 42] and content analysis of surveillance videos [45].

Due to the availability of visual and auditory information, we chose to perform the proposed techniques on theatrical movies, which contain recordings of social happenings and interactions. In order to address challenges introduced by the generality of relations among movie actors, our approach first aligns the movie script with the frames in the video using closed captions. We note that, the movie script is used only to segment the movie into scenes and provide a basis for generating the scene-actor relation matrix. Alternatively, this information can be obtained using video segmentation and face detection and recognition techniques [4, 46]. A unique characteristic of our proposed framework is its applicability to an adversarial social network, which is a highly recognized but less researched topic in sociology [39], possibly due to the complexity of defining adversarial relations alongside friendship relations. Without loss of generality, an adversarial social network contains two disjoint rival communities composed of actors, where members within a community have friendly relations and across communities have adversarial relations.

In our basic framework, we use visual and auditory information to quantify a grouping cue at the scene level, which serves as soft constraints among the actors. These soft constraints are then integrated to learn interactor affinity. The communities in the resulting social network are discovered by subjecting the interactor affinity matrix to a generalized modularity principle [5]. Each community in a social network typically contains an influential person who strongly connects to all other members of the community. Arguably, we call this influential person as the leader of the community, and detect him/her by adopting eigenvector centrality [33]. An illustration of the communities and their leaders discovered by our approach is given in Fig. 2.1 for the movie titled *G.I. Joe: The Rise of Cobra* (2009).

The remainder of the chapter is organized as follows: We start with a brief survey of related work. We then describe our learning-based approach to transform low-level features into grouping cues in Sect. 2.2. Section 2.3 details how we learn social networks from videos. The methodology used to analyze these social networks is described in Sect. 2.4, and is evaluated on a set of videos in Sect. 2.5. Furthermore, the connections between visual concepts and social relations are explored in Sect. 2.6.

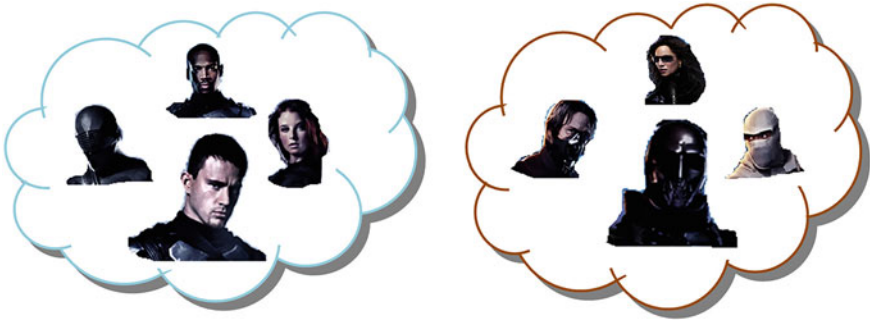
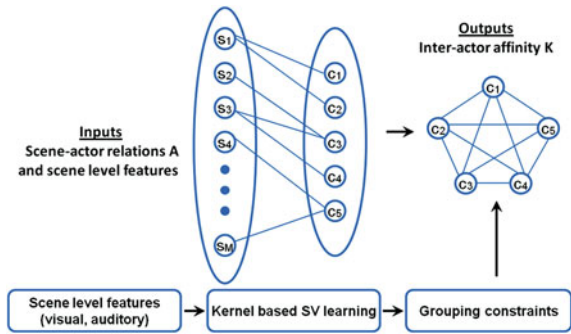


Fig. 2.1 Pictorial representation of communities in the movie titled *G.I. Joe: The Rise of Cobra* (2009). Our approach automatically detects the two rival communities (*G.I. Joe* and *Cobra*), and identifies their leaders (*Duke* and *McCullen*) visualized as the upscaled characters at the front of each community

Fig. 2.2 Flow diagram of the proposed learning-based framework for constructing social networks



We summarize the chapter in Sect. 2.7. The learning-based framework for constructing social networks is illustrated in Fig. 2.2, for an example video of M scenes and 5 interacting actors. The scene–actor relations are visualized in the top row, where a scene is linked to actors that appear in it. A social network representing the video is learned from both scene–actor relations and grouping constraints.

2.1.1 Related Work

Aside from research conducted on social networking in the field of sociology, other fields, which include data mining, computer vision, and multimedia analysis, have been using the ideology behind social networking to solve problems in their respective domains [40, 42, 45]. In this section, we expand our discussion on approaches as they relate to our problem domain.

For detecting communities from surveillance video, a recent study reported in [45] takes advantage of traditional social networking methods. The interactions between

the objects are conjectured to occur depending on the proximity heuristic. This heuristic, while far from representing social relations, defines a measurable quantity to define communities in the scene. The authors use traditional modularity [28] to find such communities. Similarly, Ge et al. [18] define the existence of social relations based on the proximity and relative velocity between the objects, which are later used to detect communities in a crowd by means of clustering techniques.

For analysis and segmentation of movies, Weng et al. [40] generate a social network from actor co-occurrence in a scene without attributing them to low-level video features. In their work, the relations are conjectured to be only friendly relations; hence, community can be detected using traditional clustering techniques. While the co-occurrence reasoning resembles our approach, we go one step further and relate them to audiovisual features which are used in a learning framework to define the types of interactions.

Besides, the latest work in computer vision including [15] has strived to identify specific categories of social interaction from video content using Markov random fields (MRFs), but the goal there is not to estimate the social network structure for a group of individuals. Similarly, the study reported in [31] deals with the problem of recognizing social roles played by people in an event via weakly supervised conditional random fields (CRFs). However, the authors did not leverage social network structure for such analysis.

Social network analytics has recently been considered in data mining to find certain interactions within a network generated from log-entries. In two different studies [14, 42] by different groups, the authors have analyzed social networks and their dynamics using Bayesian modeling of social networks and people's interactions. Similar to data mining, in reality mining, researchers have used mobile phone usage to infer social networks using nonvideo data [11, 12]. Due to ubiquitous availability of videos, we believe our framework opens up new directions for studying social phenomena from videos. While it is relatively straightforward to use log-entries or cell phone usage, extracting social content from video presents significant challenges to pattern recognition research.

Broadly speaking, the main difference of our approach from all the papers cited above and many others in the field of sociology is the methodology we have taken to address the social network generation problem. Particularly, existing methods define a heuristic interaction and derive a social network using these heuristics. Before we continue, we would like to draw an analogy between our treatment and the way humans would approach to the same problem. Consider a scene where there are several individuals performing some activities; at first sight, a human observer without knowing the domain and interaction types would immediately consider all individuals are equally related to each other. Based on the duration the observer watches the scene, he will start guessing the type of interactions and derive a social network using his past experience. Similar to human experience, we learn interactions and networks from given training examples and infer social networks in a novel scene. We believe, the proposed approach can benefit other areas in computer vision such as meeting video analysis where modeling individual actions and high-level relations between

Table 2.1 A summary of related work in addition to ours on constructing social networks from data

Data sources	Observed features	Construction techniques	Usability of framework	Examples
From interaction logs	Social interactions	Simple connections	On collected social data, e.g. emails	[14, 42]
From cell phone usage	Call data, etc.	Simple connections	On collections of mobile devices	[11, 12]
From videos (existing)	Tracked people	Proximity heuristics	On surveillance videos	[18, 45]
From videos (this chapter)	Audio-visual cues	Learning approaches	On videos with training labels	[8, 9]

the attendees are important [1, 47]. Finally, some of the aforementioned studies are summarized in Table 2.1, where the novelty and generality of our framework can be readily observed.

2.2 Learning Grouping Cues

Consider a case when the relations among actors have no prior specification and what we observe is only low-level video features. In this setting, communities in the video cannot be explicitly labeled. In our approach, we extract the low-level features from videos, such that kernels on scene-level features provide grouping cues using regression learned from other videos. Unlike many existing approaches, the proposed mapping strategy is data-driven and provides a flexible and extensible approach that can incrementally use new features as they become available.

Before we proceed, let us assume that a video \mathbb{V} is composed of scenes, $s_1, s_2 \dots s_M$, each of which contains a set of actors and has an associated grouping cue $\beta_i \in [-1, +1]$. The grouping cue serves as a basis to decide whether actors co-occurring in the scene belong to the same community ($\beta_i > 0$) or different communities ($\beta_i < 0$). In our setting, the larger the absolute value of β_i is, the more stringent the corresponding constraints are. In the following discussion, we will detail our approach on estimating such grouping cues from low-level video content.

We conjecture that the interactions among the members of a social community are different from the interactions among the members across different communities. This conjecture imposes a weak grouping assumption due to the fact that we do not need to know the identities of interacting communities. Therefore, labeling from a set of source videos can be generalized to a novel video. In a similar manner, we also conjecture that a video of activities contains low-level features that convey the characteristic types of relationships among the actors performing them. In other words, the relationships between the activities and the actors provide a distinct feature set

that can be used to infer if members of a single community or different communities co-occur in the same video segment. For example, boys and girls attending a school interact in distinct ways within and across the groups [19, 27]. Similarly, rivalry across different groups creates adversarial relations when they interact, such as the actions they perform and the words they exchange.

Considering that a scene is the smallest segment in a movie which contains a continued event, low-level features generated from the video and audio of each scene can be used to quantify adversarial and non-adversarial contents. Movie directors often follow certain rules, referred to as the film grammar or cinematic principles in the film literature, to emphasize the adversarial content in scenes. Typically, adversarial scenes contain abrupt changes in visual and auditory contents, whereas these contents change gradually in non-adversarial scenes. Therefore, the visual and auditory features, which quantify adversarial scene content, can be extracted by analyzing the disturbances in the video [32].

In particular for measuring visual disturbance, we follow the cinematic principles and conjecture that for an adversarial scene, the motion field is nearly evenly distributed in all directions (see Fig. 2.3 for illustration). For generating the optical flow distributions, we use the Kanade–Lucas–Tomasi tracker [34] within the scene bounds and use good features to track. Alternatively, one can use dense flow field generated by estimating optical flow at each pixel [26]. The visual disturbance in the observed flow field can be measured by entropy of the orientation distribution as shown in Fig. 2.4. Specifically, we apply a moving window of 10 frames with 5 frames overlapping in the video for constructing the orientation histograms of optical flows. We use histograms of optical flow vectors weighted by the magnitude of motion. The number of orientation bins is set to 10 and the number of entropy bins in the final feature vector is set to 5. As can be observed in Fig. 2.5, flow distributions generated from adversarial scenes tend to be uniformly distributed and thus, they consistently have more high-entropy peaks compared to non-adversarial scenes. This observation serves as the basis for distinguishing the two types of scenes.

Auditory features extracted from the accompanying movie audio are used together with the visual features to improve the performance. We adopt a combination of temporal and spectral auditory features discussed in [24, 32], which are energy peak ratio, energy entropy, short-time energy, spectral flux, and zero crossing rate:

- Energy peak ratio $EPR = \frac{p}{S}$, where p is the number of energy peaks and S is length of an audio frame;
- Energy entropy $EE = -\sum_{i=1}^K e_i \log e_i$, where a audio frame is divided into K sub-windows. For sub-window i , energy e_i is computed;
- Short-time energy $SE = \sum_{i=1}^S x_i^2$, where S is the length of an audio frame;
- Spectral flux $SF = \frac{1}{KF} \sum_{i=2}^K \sum_{j=1}^F (\varepsilon_{i,j} - \varepsilon_{i-1,j})^2$, where $\varepsilon_{i,j}$ is the spectral energy at sub-window i and frequency channel j ;
- Zero crossing rate $ZCR = \frac{1}{2S} \sum_{i=1}^S |sgn(x_i) - sgn(x_{i-1})|$, where sgn stands for a sign function.

Specifically, these features are computed for sliding audio frames that are 400 ms in length. The means of these features over the duration of the scene constitute

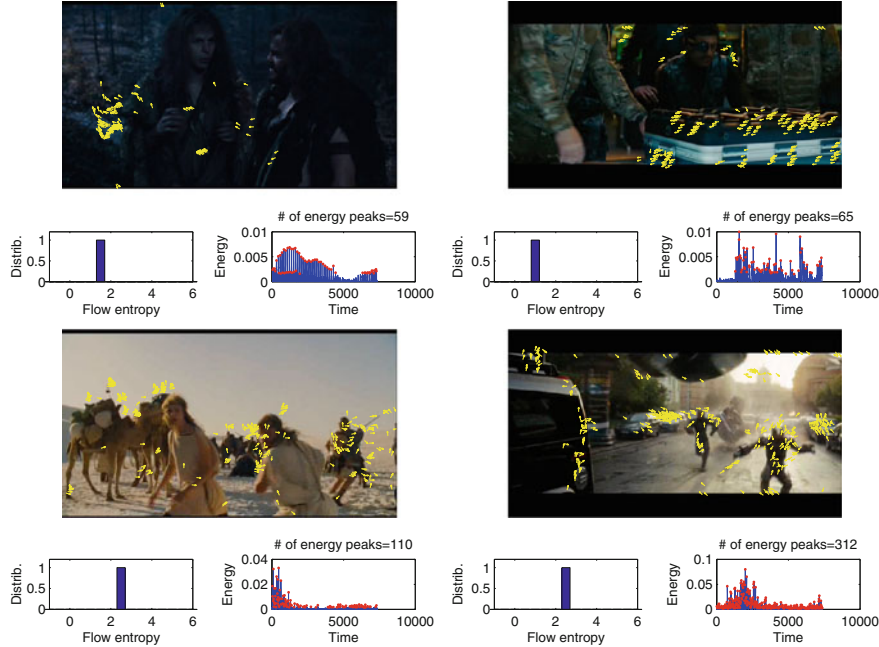


Fig. 2.3 Visual and auditory characteristics of adversarial scenes. *Top row* non-adversarial scenes from *Year One* (2009) and *G.I. Joe: The Rise of Cobra* (2009); *Bottom row* adversarial scenes from these two movies. Optical flow vectors are superimposed on the frames and computed features are shown as plots for a temporal window of 10 video frames, including entropy distribution of optical flow vectors and detected energy peaks (red dots in energy signals). *Note* For color interpretation see online version

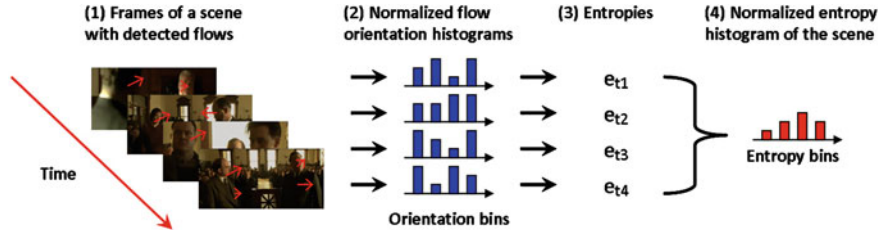


Fig. 2.4 Generation of the normalized entropy histogram from orientation distributions of optical flows detected from a scene

a feature vector. A sample auditory feature (energy peaks) is shown in Fig. 2.3 for both adversarial and non-adversarial scenes. It can be observed that adversarial scenes have more peaks in energy signals, which are moving averages of squared audio signals.

The visual and auditory features provide two vectors per scene (5 dimensional visual and 5 dimensional auditory), which are used to estimate a real-valued group-

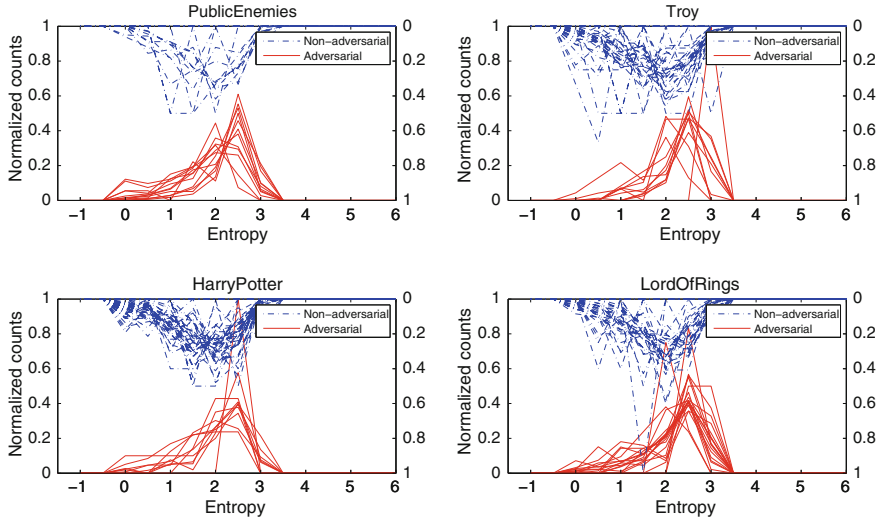


Fig. 2.5 Visualization of entropy histogram feature vectors extracted from four example movies. The two classes (adversarial and non-adversarial) have distinct patterns, in that adversarial scenes tend to consistently produce strong peaks in high entropies. Best viewed in color

ing cue $\beta_i \in [-1, +1]$ of the scene s_i . To achieve this goal, we use support vector regression (SVR) [35], which has been successfully used to solve various problems in computer vision literature [10, 21]. We apply a radial basis function to both the visual and auditory feature vectors, which leads to two kernel matrices \mathcal{K}_v and \mathcal{K}_a , respectively. The two kernel bandwidths can be chosen by using cross-validation. The joint kernel is then computed as the multiplication kernel: $\hat{\mathcal{K}}(u, v) = \mathcal{K}_v(u, v)\mathcal{K}_a(u, v)$. In support vector regression, 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 that the final decision function can be written as:

$$\beta_i = g(s_i) = \sum_{j=1}^L (\alpha_j - \alpha_j^*) \hat{\mathcal{K}}_{l_j, i} + b, \quad (2.1)$$

where the coefficient b is offset, α_i and α_i^* are the Lagrange multipliers for labeling constraints, L is the number of labeled examples, and l_j is the index for the j th labeled example.

In our problem domain, the joint kernel together with training video scenes and their grouping cues $\beta_i = +1$ (scene with members of only one 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 grouping cues β_i using the regression learned from labeled video scene examples from other videos in the training set.

2.3 Learning Social Networks

Consider actors co-occurring in a video. We conjecture that these actors co-occur more often if they are members of the same community. The higher the number of co-occurrences for the same-community members is, the more positive grouping cues are present compared to the negative ones. The combination of these two factors plays a significant role in our social network learning methodology. In the following discussion, we will first describe the representations we use which is followed by how social relations are learned.

2.3.1 Basic Representations

The temporal occurrence of an actor c_i in a video is represented by a boolean appearance function $\psi_i : T \rightarrow \{0, 1\}$, where the duration of a video $T \subset \mathbb{R}^+$. In practice, we only have access to its sampled version. Suppose that the sampling period is of length t seconds. According to the Nyquist's sampling theorem, as long as $t \leq \min_i \{1/2B_i\}$, where B_i is the highest frequency (in hertz) of actor i 's appearance function, the continuous appearance information and their co-occurrence can be determined from those discrete samples.

A video \mathbb{V} is composed of nonoverlapping M scenes, where each scene s_i contains interactions among a set of actors. In order for computational convenience, the appearance functions of actors are approximated as a scene-actor relation matrix denoted by $A = \{A_{i,j}\}$, where $A_{i,j} = 1$ if there exists $t \in L_i$, where L_i is the temporal interval of s_i , such that $\psi_j(t) = 1$. This, for a movie, can be obtained by searching for speaker names in the script. This representation is reminiscent of the actor-event graph in social network analysis [39]. While the actor relations in A can be directly used for construction of the social network, we will demonstrate later that the use of visual and auditory scene features can lead to a better social network representation.

Finally, the social network is represented as an undirected graph $G(V, E)$ with cardinality $|V|$. In this graph, the vertices represent the actors

$$V = \{v_i : \text{vertex } v_i \sim \text{actor } c_i\}, \quad (2.2)$$

and the edges define the interactions between the actors

$$E = \{(v_i, v_j) | v_i, v_j \in V\}. \quad (2.3)$$

The resulting graph G is a fully connected graph with an affinity matrix K of size $|V| \times |V|$. The entry in the affinity matrix $K(c_i, c_j)$ for two actors c_i and c_j is a real-valued weight, which is decided by an affinity learning method that will be

introduced next in this section. The values in the affinity matrix serve as the basis for social network analytics, including determining communities and leaders in the social network.

2.3.2 Actor Interaction Model

Let c_i be actor i , and $\mathbf{f} = (f_1, \dots, f_N)^T$ be the vector of community memberships containing ± 1 values, where f_i refers to the membership of c_i . Let \mathbf{f} distribute according to a zero-mean identity-covariance Gaussian process

$$P(\mathbf{f}) = (2\pi)^{-N/2} \exp^{-\frac{1}{2} \mathbf{f}^T \mathbf{f}}. \quad (2.4)$$

In order to model the information contained in the scene–actor relation matrix A and the aforementioned grouping cue of each scene β_i , we assume the following distributions:

- if c_i and c_j occur in a non-adversarial scene k ($\beta_k \geq 0$), we assume $f_i - f_j \sim \mathcal{N}(0, \frac{1}{\beta_k^2})$;
- if c_i and c_j occur in an adversarial scene k ($\beta_k < 0$), we assume $f_i + f_j \sim \mathcal{N}(0, \frac{1}{\beta_k^2})$.

Therefore, if $\beta_i = 0$, then the constraint imposed by a scene becomes inconsequential, which corresponds to the least confidence in the constraint. On the other hand, if $\beta_i = \pm 1$, the corresponding constraint becomes the strongest. Because of the distributions we use, none of the constraints is hard, making our method relatively flexible and insensitive to prediction errors. Applying the Bayes' rule, the posterior probability of \mathbf{f} given the constraints is defined in a continuous formulation as the following:

$$\begin{aligned} P(\mathbf{f} | \{\psi_k\}, \beta) &= P(\mathbf{f}) \exp \left\{ - \sum_{i,j} \int_{t \in \{t: \beta(t) \geq 0\}} \psi_i(t) \psi_j(t) \frac{(f_i - f_j)^2 \beta(t)^2}{2} dt \right. \\ &\quad \left. - \sum_{i,j} \int_{t \in \{t: \beta(t) < 0\}} \psi_i(t) \psi_j(t) \frac{(f_i + f_j)^2 \beta(t)^2}{2} dt \right\} \\ &\propto \exp \left\{ - \frac{1}{2} \mathbf{f}^T \mathbf{f} - \sum_{i,j} \int_{t \in \{t: \beta(t) \geq 0\}} \psi_i(t) \psi_j(t) \frac{(f_i - f_j)^2 \beta(t)^2}{2} dt \right. \\ &\quad \left. - \sum_{i,j} \int_{t \in \{t: \beta(t) < 0\}} \psi_i(t) \psi_j(t) \frac{(f_i + f_j)^2 \beta(t)^2}{2} dt \right\}. \end{aligned} \quad (2.5)$$

Translating this equation into its discrete version, we have:

$$\begin{aligned}
 P(\mathbf{f}|A, \beta) &= P(\mathbf{f}) \prod_{k:\beta_k \geq 0} \prod_{c_i, c_j \in s_k} \exp \frac{-(f_i - f_j)^2 \beta_k^2}{2} \\
 &\quad \prod_{k:\beta_k < 0} \prod_{c_i, c_j \in s_k} \exp \frac{-(f_i + f_j)^2 \beta_k^2}{2} \\
 &\propto \exp \left\{ -\frac{1}{2} \mathbf{f}^T \mathbf{f} - \sum_{k:\beta_k \geq 0} \sum_{c_i, c_j \in s_k} \frac{(f_i - f_j)^2 \beta_k^2}{2} \right. \\
 &\quad \left. - \sum_{k:\beta_k < 0} \sum_{c_i, c_j \in s_k} \frac{(f_i + f_j)^2 \beta_k^2}{2} \right\}. \tag{2.6}
 \end{aligned}$$

It can be verified that $P(\mathbf{f}|A, \beta) \propto \exp(-\frac{1}{2} \mathbf{f}^T K^{-1} \mathbf{f})$ is a Gaussian process with zero mean. Using $K_{i,j} = E\{f_i f_j | A, \beta\}$ as the learned affinity between c_i and c_j , it follows that $K = M^{-1}$, where

$$M_{i,j} = \begin{cases} \sum_{k:c_i, c_j \in s_k, \beta_k < 0} \beta_k^2 - \sum_{k:c_i, c_j \in s_k, \beta_k \geq 0} \beta_k^2 & \text{if } i \neq j \\ 1 + \sum_{l \neq i} \sum_{k:c_l, c_i \in s_k} \beta_k^2 & \text{if } i = j \end{cases}. \tag{2.7}$$

The resulting K is symmetric and positive definite. However, unlike an affinity matrix from a Gaussian kernel, it may contain negative values. The proposed approach has two special cases:

- In the case when $\beta_i = 1$, then the aforementioned learning mechanism reduces to a co-occurrence-based approach which is a traditional tool in social network analysis [5, 28]. Specifically, $M_{i,j}$, for $i \neq j$, represents the minus value of the number of scenes where c_i and c_j occur together. This reduced scheme does not utilize the video/audio feature-based prediction of grouping cues, and serves as a natural baseline in this chapter.
- If we use fixed variance parameters in the assumed distributions instead of the learned ones, our affinity learning method reduces to the affinity propagation approach proposed in [25].

2.4 Social Network Analytics

A primary goal of social network analytics is finding groups of actors to form social communities and detecting the most influential actor within a community, which we arguably refer to as the leader of a community. Traditionally, communities are detected using spectral clustering techniques tailor-made for social settings, such as the popular modularity-cut algorithm [28]. A recent study reported in [5] has shown

that the performance of modularity cut can be increased by introducing a generalized objective referred to as the max–min modularity. The max–min modularity clustering, however, assumes unweighted edges and is not directly suitable for our social networks which contain learned weighted edges.

In our design, we first generate a principal affinity matrix K' by the following rules: $K'_{i,j} = K_{i,j}$ for $K_{i,j} > 0$, and $K'_{i,j} = 0$ for other entries. We then generate a complementary affinity matrix K'' by the following rules: $K''_{i,j} = -K_{i,j}$ for $K_{i,j} < 0$, and $K''_{i,j} = 0$ for other entries. The matrix K'' represents the unrelatedness between vertices in the network in terms of community memberships. Adopting the strategy in [5] and using K' and K'' , we formulate the max–min modularity criterion as $Q_{MM} = Q_{\max} - Q_{\min}$ for:

$$Q_{\max} = \frac{1}{2m'} \sum_{i,j} \left(K'_{ij} - \frac{k'_i k'_j}{2m'} \right) (f_i f_j + 1) \triangleq \frac{1}{2m'} \sum_{i,j} B'_{i,j} (f_i f_j + 1), \quad (2.8)$$

$$Q_{\min} = \frac{1}{2m''} \sum_{i,j} \left(K''_{ij} - \frac{k''_i k''_j}{2m''} \right) (f_i f_j + 1) \triangleq \frac{1}{2m''} \sum_{i,j} B''_{i,j} (f_i f_j + 1), \quad (2.9)$$

where $m' = \frac{1}{2} \sum_{i,j} K'_{ij}$, $k'_i = \sum_j K'_{ij}$, $m'' = \frac{1}{2} \sum_{i,j} K''_{ij}$, $k''_i = \sum_j K''_{ij}$ and the term $\frac{k'_i k'_j}{2m'}$ represents the expected edge strength between the actors c_i and c_j [28]. Based on this observation, we note that $K'_{i,j} - \frac{k'_i k'_j}{2m'}$ measures how much the connection between two actors is stronger than what would be expected between them, and serves as the basis for keeping the two actors in the same community. In this formulation, the max–min modularity Q_{MM} roots from the conditions for a good network division that (1) edge strength across communities should be smaller than expected, and (2) unrelated actors within a community should be minimal. These conditions can be realized by maximizing Q_{MM} . Using standard eigenanalysis, it follows that the eigenvector \mathbf{u} of $\frac{1}{2m'} B' - \frac{1}{2m''} B''$ with the largest eigenvalue maximizes a relaxed version of Q_{MM} . The resulting eigenvector solution contains real values, and we threshold them at the 0 level to obtain the desired community memberships. That is, we let $f_i = +1$ if $u_i \geq 0$, and $f_i = -1$ if otherwise.

Once the communities in the video are extracted, their leaders are detected by analyzing the centrality of each actor in the community. In the sociology literature, the centrality of an actor is traditionally defined by its degree or betweenness [17]. In this chapter, we, rather, adopt a new technique which is referred to as the eigen-centrality [33] due to its relation to proposed community detection approach. Let the centrality score, x_i for the i th actor be proportional to the sum of the scores of all vertices which are connected to it: $x_i = \frac{1}{\lambda} \sum_{j=1}^N K'_{i,j} x_j$, where N is the total number of actors in the video and λ is a constant. It follows from this notation that the centralities of actors satisfy $K' \mathbf{x} = \lambda \mathbf{x}$ in the vector form. It can be shown that the eigenvector with largest eigenvalue provides the desired centrality measure [33]. Therefore, if we let the eigenvector of K' with the largest eigenvalue be \mathbf{v} , the leaders of the two

communities are given by $\arg \max_{i:u_i \geq 0} v_i$ and $\arg \max_{i:u_i < 0} v_i$, respectively. In our problem domain, when the communities correspond to two adversarial social groups, their expected leaders relate to the hero or the villain in the video.

2.5 Experiments

For qualitative and quantitative evaluation of the proposed approach, we generate a dataset of 10 movies which contains recent and classical theatrical movies that cover a range of genres including action, adventure, fantasy, and drama.¹ The movies in our dataset broadly contain two rival communities with a designated leader for each community. For each movie with statistics tabulated in Table 2.2, the dataset contains visual and auditory features, movie script, and closed caption data, all of which are temporally aligned.

For movie domain, in order to align visual and auditory features with the script, we require temporal segmentation of the movie into scenes, which provides start and stop timings for each scene. This segmentation process is guided by the accompanying movie script and closed captions. The script is usually a draft version with no time tagging and lacks professional editing, while the closed captions are composed of lines d_i , which contain timed sentences uttered by actors. The approach we use to perform this task can be considered as a variant of the alignment technique in [6]:

1. Divide the script into scenes, each of which is denoted as s_i . Similarly, closed captions are divided into lines d_i .
2. Define \mathcal{C} to be a cost matrix. Compute the percentage p of the words in closed caption d_j matched with scene s_i while respecting the order of words. Set the cost as $\mathcal{C}_{i,j} = 1 - p$.
3. Apply dynamic time warping to \mathcal{C} for estimating start t_1^i and stop times t_2^i of s_i , which respectively correspond to the smallest and largest time stamps for closed captions matched with s_i .

At the end of this process, we generate the start and stop timings of all scenes. Due to the fact that publicly available scripts for movies are not perfectly edited, the temporal segmentation may not be precise. Nevertheless, our approach is robust to such inaccuracies in segment boundaries.

In the following discussion, we analyze social networks with accompanying affinity matrices generated from





















- the actor co-occurrence information reflected in matrix A (co-occurrence), which is more traditional in sociology;

¹ The movies in our dataset are (1) *G.I. Joe: The Rise of Cobra* (2009); (2) *Harry Potter and the Half-Blood Prince* (2009); (3) *Public Enemies* (2009); (4) *Troy* (2004); (5) *Braveheart* (1995); (6) *Year One* (2009); (7) *Coraline* (2009); (8) *True Lies* (1994); (9) *The Chronicles of Narnia: The Lion, the Witch and the Wardrobe* (2005); and (10) *The Lord of the Rings: The Return of the King* (2003).

Table 2.2 Statistics of movies in our dataset which includes the number of scenes in the movie, the number of lines in closed caption data, the total number of actors in the movie, and the number of actors in one of the two communities

Movies enumerated in footnote 1	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
# of scenes	198	151	238	226	116	51	105	297	188	199
# of captions in lines	1,143	1,585	1,063	1,155	1,337	1,515	1,293	1,262	1,099	1,402
# of actors in total	11	7	10	10	7	7	6	8	10	9
# of actors in community 1	6	4	6	5	3	3	4	6	7	7

Table 2.3 Community leaders discovered using the proposed framework

Movies	(1)	(2)	(3)	(4)	(5)
Community 1	 <i>Hawk</i>	 Harry	 Dillinger	 <i>Achilles</i>	 <i>MacClan.</i>
Community 2	 McCullen	 Snape	 Purvis	 <i>Androm.</i>	 Longsha.
Movies	(6)	(7)	(8)	(9)	(10)
Community 1	 Zed	 Coraline	 <i>Trilby</i>	 <i>Susan</i>	 Frodo
Community 2	 <i>Abraham</i>	 OtherMo.	 <i>Salim</i>	 <i>Witch</i>	 WitchKing

The names in bold face refer to correct ones, whereas those in italics are not

- in addition to co-occurrence, scene-level grouping cues β_i which are learned from video and audio contents using the proposed approach.

In order to evaluate the contribution of these features, we provide comparisons of collective use of visual and auditory features with their individual use in extraction of communities and their leaders. In Fig. 2.6, we show graphical representations of the social networks for 10 movies learned from both visual and auditory features using the proposed approach. The color codes in the figure reflect the strength of affinity between actors. We observe that intercommunity connections tend to be weaker than certain intracommunity ones.

The affinity between the actors is strongly related to the grouping cues of the scenes in which they appear. This relation suggests validation of how effective the support vector regression (SVR) is for estimating the grouping cues. In order to facilitate this, we compute the mean square error (MSE) as our error measure over all the scenes in each movie, and average the resulting MSEs over all 10 movies. When both the visual and auditory features are used, the MSE is estimated as 0.61. In contrast, when only one of the features is used MSE increases to 0.80 for visual only and 0.77 for auditory only. These numbers translate into accuracy rates for

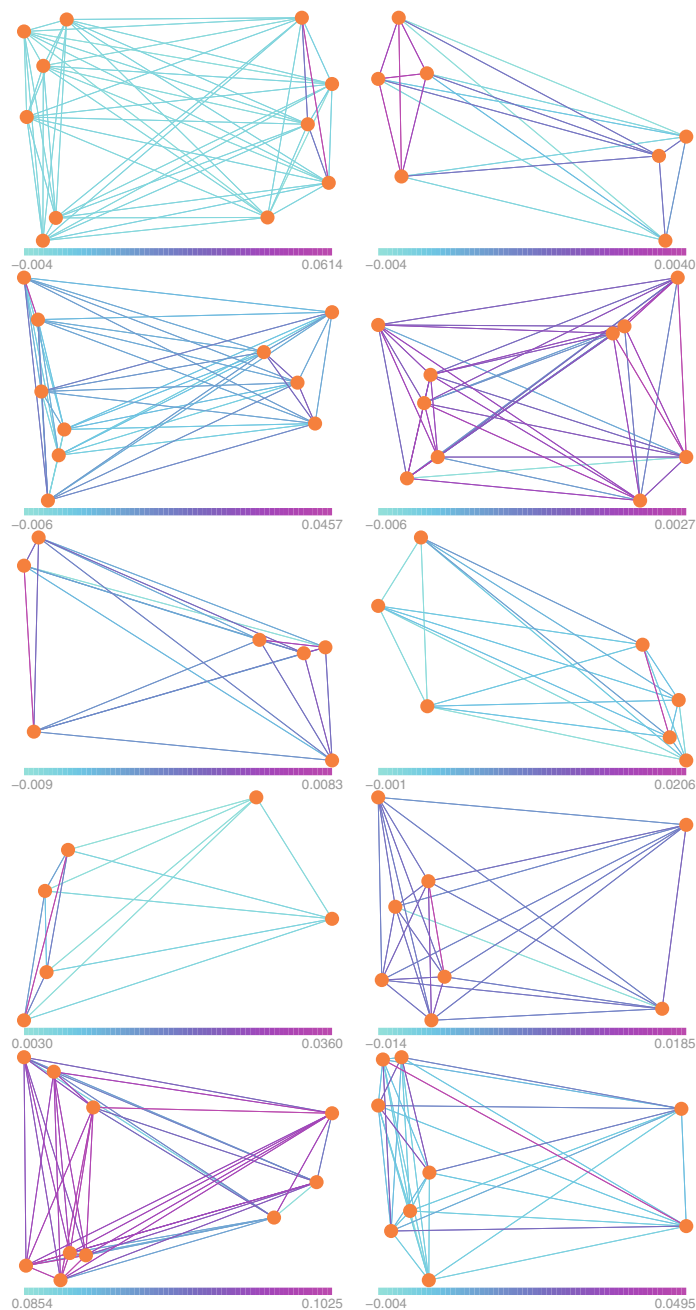


Fig. 2.6 Social networks generated using the proposed approach for the 10 movies in our dataset. Actors (vertices) are placed on the *left* and *right* with respect to communities they belong to. The strength of affinity is indicated by *pinkness* of the edges: the stronger the edge is the *pinkier* it is. Best viewed in color. *Note* For color interpretation see online version

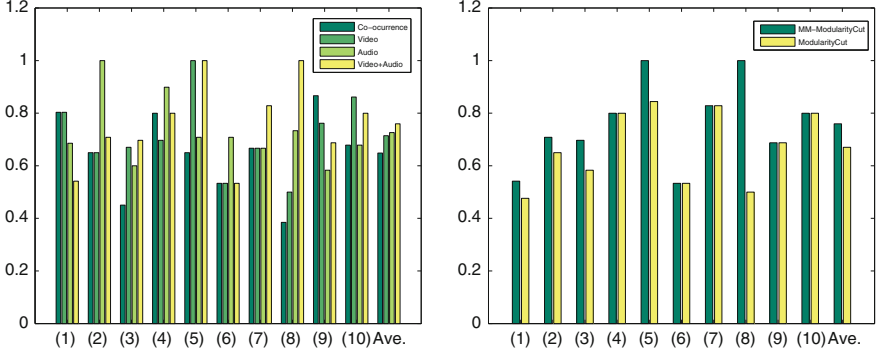


Fig. 2.7 Accuracy of social network analytics in F_1 measures. *Left*: comparison of four approaches, where the proposed one is video+audio; *Right*: comparison of two modularity algorithms (max-min vs. original), with the proposed video+audio approach

predicting if a scene is adversarial or non-adversarial. Respectively, the accuracy rates are computed as 81.6, 78.2, and 78.7% for collective feature use, visual only and auditory only. These numbers reflect that the grouping cue estimates of scenes can be further utilized to infer the relations among the movie actors.

The accuracy of community detection relates to how precise the assignment of the actors is into each one of the community. Considering that a community is a set of actors, the accuracy can be measured using the precision and recall values of predicted assignments given the ground truth. For each community these two values can be combined into an F_1 measure, which is the harmonic mean of precision and recall. This measure takes into account the possible imbalance in the size of communities and has been widely adopted. Considering that the movies in our dataset contains more than one community, we report the average F_1 measure over detected communities as the final detection accuracy for each movie.

From the quantitative evaluations shown in Fig. 2.7, for four movies visual features help enhance performance appreciably. Overall, auditory features improve the performance slightly more than the visual features when they are used independently. Their combination, however, provides the best performance, which on the average leads to an F_1 measure of 76.0%. This score, when compared to using only the actor co-occurrence to generate the social network, improves the grouping performance by 11.1%. In the same figure, we also show that the modified max-min modularity, when compared to the traditional modularity computed from K , improves the F_1 measure by 8.9%.

As discussed in Sect. 2.4, the community assignment of actors is realized by analyzing the eigenspace of $\frac{1}{2m'}B' - \frac{1}{2m''}B''$. In order to visualize this assignment process, we map the actors in the movie into coordinates defined by the two eigenvectors with highest eigenvalues. This mapping provides an optimal way to visualize the interactor relations in two dimensions. In Fig. 2.8, we illustrate the ground truth

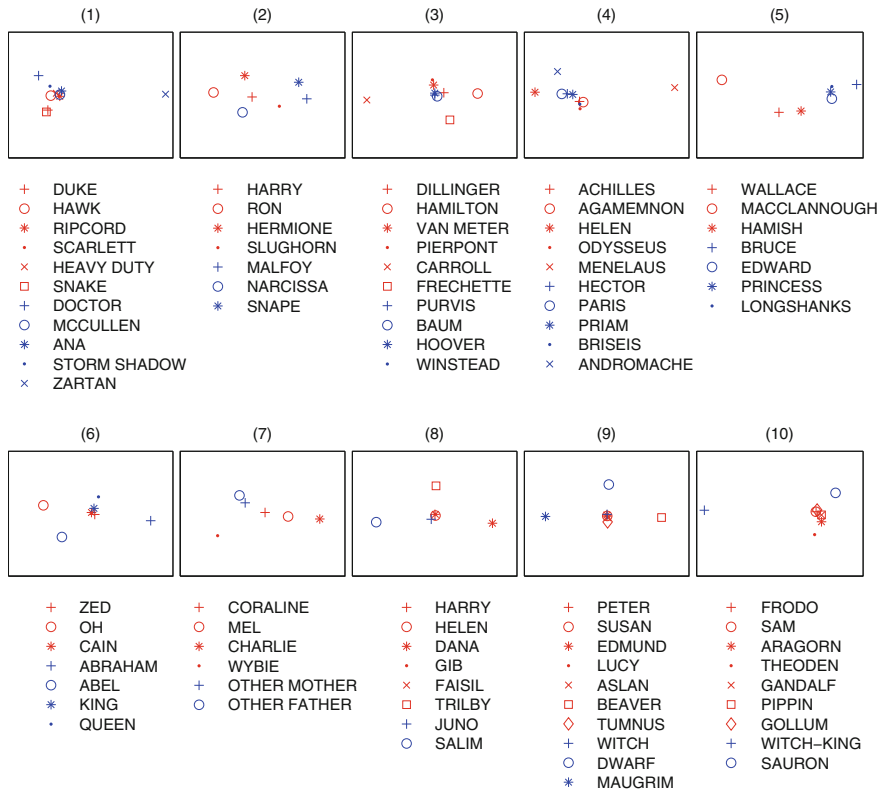


Fig. 2.8 2D visual maps of actor relations. *Red* and *blue* stand for the two communities, respectively, according to our ground truth labeling. Best viewed in color. *Note* For color interpretation see online version

in red and blue colors, respectively, for the two communities.² As can be observed, the actors who belong to separate communities tend to lie apart.

As we mentioned earlier, the eigenvector of K' with the highest eigenvalue provides the leaders of communities. In Table 2.3, we tabulate these leaders with their pictures for the two rival communities for each movie. The predicted leaders who correspond to the true leaders in the movie are shown in bold face, while incorrect leaders are shown in *italics*. Overall, it can be observed that many of the leaders are successfully discovered by our framework.³

² In movie (10), *Gollum* has a good personality except for when he is close to the ring. The ring changes the good behavior of the actors to bad except for *Frodo*.

³ Ground truth leaders are: (1) Duke and McCullen; (2) Harry and Snape; (3) Dillinger and Purvis; (4) Achilles and Hector; (5) Wallace and Longshanks; (6) Zed and King; (7) Coraline and Other Mother; (8) Harry and Salim; (9) Aslan and Witch; and (10) Frodo and Witch-king.

2.6 Using Visual Concepts

In this section, we discuss an extension to the basic framework of learning social relations from videos. Visual concepts, such as “beach”, “cheering”, and “shooting”, are detected from the video content and employed as mid-level representation, which is used as the basis for inferring social relations. The main intuition behind this approach is that visual concept detection, compared to low-level video information, provides useful semantic features for inferring social relations. For example, individuals involved in a fighting scene tend to be enemies, while individuals jogging leisurely together tend to be friends. To this end, we leverage support vector machine (SVM) detectors for 374 trained visual concepts provided in [41], due to its public availability and broadness.

For detecting base concepts, we use the three low-level features exploited in [41], which include the color, texture, and edges computed from keyframes in a video. In specific, 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.

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 keyframes, we compute low-level features for each keyframe i . These features are then used to estimate a normalized score, which is the raw decision value from an SVM transformed by a logistic function, from the SVMs which are independently trained on each low-level feature: $f_{i,j}$ for three feature types $j = 1, 2, 3$. The average of the three scores $f_i = \frac{1}{3}(f_{i,1} + f_{i,2} + f_{i,3})$ 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. Since previous research has suggested that grid- and global-based feature representations extracted from keyframes and SVM classification lead to strong concept detection systems [41], this mid-level representation is used as the basis for inferring social relations.

It is clear that not all the dimensions of the semantic vectors are equally informative toward social relations. Besides, detecting some of the visual concepts may be unsatisfactory due to their large variability or relatively small spatiotemporal extents. To address this issue, we employ a supervised dimension reduction method known as kernel local Fisher discriminant analysis (KLFDA) [37], which has an analytic form of the embedding transformation. By applying this data-dependent transform,

Table 2.4 Comparative analysis of features for social relational learning

Methods	Prec(+)	Prec(-)	Prec(ave) (%)	F_1 (%)
Baseline features	—	—	78.2	76.0
Visual concepts (d=50)	83.1 %	85.0 %	84.1	81.9

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

we derive a more informative and compact representation, which is a d -dimensional vector for each video scene, for learning the social relations.

Following Sect. 2.2, we assume that a video is composed of M scenes, $s_1, s_2 \dots s_M$, each of which contains a set of actors and has an associated grouping cue β_i . The grouping cue serves as a basis to decide whether the actors co-occurring in a scene belong to the same or different communities. To estimate the grouping cues β_i from the d -dimensional transformed semantic vectors, we use support vector regression, with a radial basis function kernel over the d -dimensional transformed concept score vectors. On the same 10-movie dataset, we are able to compare the performance of using visual concepts versus the other approach in estimating grouping cues. Quantitative analysis tabulated in Table 2.4 is performed on the baseline features detailed in Sect. 2.2 versus the visual concepts described in this section. It should be noted that the F_1 score associated with visual concepts is partially accounted for by the community detection method in [9]. However, it can be seen that visual concept-based features work better than the baseline features by a large margin in the crucial task of grouping cue estimation, with no use of tailored visual features toward the “adverseness” of a scene. Besides, the generality of visual concepts for social relational learning is confirmed by a larger scale study, such as that in [9].

2.7 Summary

In this chapter, we have presented a framework for learning the relations among actors from videos using a social network approach. We have used visual and auditory features to characterize the grouping cues. By using an affinity learning procedure, we incorporate these grouping cues, and make informed decisions in constructing and analyzing the corresponding social network. Extensive analysis on a set of videos has validated the effectiveness of our framework in high-level understanding of social interactions. Besides, leveraging visual concepts has also been considered as mid-level representation for inferring social relations.

Further, it is natural to apply our framework to other problem domains, such as surveillance videos, where behaviors of interest can be related to the interactions between objects in a scene, and meeting videos, where people’s modes of interaction may be related to their social groups or organizations. In such settings, it may be

possible to include recognizable human actions in generating grouping cues. The proposed framework also contributes to sociology, in that our framework can aid sociological discovery by automatically extracting communities from videos, given available group labeling of the desired type. Uncovering and quantifying such patterns of generalizability may be of interest to sociologists.

References

1. Al-Hames, M., Lenz, C., Reiter, S., Schenk, J., Wallhoff, F., Rigoll, G.: Robust multi-modal group action recognition in meetings from disturbed videos with the asynchronous hidden markov model. In: International Conference on Image Processing (2007)
2. Ali, S., Basharat, A., Shah, M.: Chaotic invariants for human action recognition. In: IEEE International Conference on Computer Vision (2007)
3. Alon, J., Athitsos, V., Yuan, Q., Sclaroff, S.: A unified framework for gesture recognition and spatiotemporal gesture segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(9), 1685–1699 (2009)
4. Arandjelović, O., Zisserman, A.: Automatic face recognition for film character retrieval in feature-length films. In: ACM International Conference on Image and Video Retrieval (2005)
5. Chen, J., Zaiane, O., Goebel, R.: Detecting communities in social networks using max-min modularity. In: SIAM Conference on Data Mining (2009)
6. Cour, T., Jordan, C., Miltsakaki, E., Taskar, B.: Movie/script: alignment and parsing of video and text transcription. In: European Conference on Computer Vision (2008)
7. Ding, L., Fan, Q., Hsiao, J., Pankanti, S.: Graph based event detection from realistic videos using weak feature correspondence. In: International Conference on Acoustics, Speech, and Signal Processing (2010)
8. Ding, L., Yilmaz, A.: Learning relations among movie characters: a social network perspective. In: European Conference on Computer Vision (2010)
9. Ding, L., Yilmaz, A.: Inferring social relations from visual concepts. In: International Conference on Computer Vision (2011)
10. Duffrenois, F., Colliez, J., Hamad, D.: Crisp weighted support vector regression for robust single model estimation: application to object tracking in image sequences. In: IEEE Conference on Computer Vision and Pattern Recognition (2007)
11. Eagle, N., Pentland, A.: Eigenbehaviors: identifying structure in routine. *Behav. Ecol. Sociobiol.* **63**(7), 1057–1066 (2009)
12. Eagle, N., Pentland, A., Lazer, D.: Inferring social network structure using mobile phone data. *Proc. Nat. Acad. Sci.* **106**(36), 15274–15278 (2009)
13. Efros, A.A., Berg, A.C., Mori, G., Malik, J.: Recognizing action at a distance. In: IEEE International Conference on Computer Vision (2003)
14. Fan, Y., Shelton, C.R.: Learning continuous-time social network dynamics. In: Conference on Uncertainty in Artificial Intelligence (2009)
15. Fathi, A., Hodgins, J.K., Rehg, J.M.: Social interactions: a first-person perspective. In: IEEE Conference on Computer Vision and Pattern Recognition (2012)
16. Fathi, A., Mori, G.: Action recognition by learning mid-level motion features. In: IEEE Conference on Computer Vision and Pattern Recognition (2008)
17. Freeman, L.: Centrality in social networks: conceptual clarification. *Soc. Netw.* **1**(3), 215–239 (1979)
18. Ge, W., Collins, R., Ruback, B.: Automatically detecting the small group structure of a crowd. In: IEEE Workshop on Applications of Computer Vision (2009)
19. Holden, C.: Giving girls a chance: patterns of talk in co-operative group work. *Gend. Educ.* **5**(2), 179–189 (1993)

20. Jiang, H., Fels, S., Little, H.: A linear programming approach for multiple object tracking. In: IEEE Conference on Computer Vision and Pattern Recognition (2007)
21. Kusakunniran, W., Wu, Q., Zhang, J., Li, H.: Support vector regression for multi-view gait recognition based on local motion feature selection. In: IEEE Conference on Computer Vision and Pattern Recognition (2010)
22. Kyriazis, N., Argyros, A.: Physically plausible 3d scene tracking: the single actor hypothesis. In: IEEE Conference on Computer Vision and Pattern Recognition (2013)
23. Laptev, I., Lindeberg, T.: Space-time interest points. In: IEEE International Conference on Computer Vision (2003)
24. Lin, J., Wang, W.: Weakly-supervised violence detection in movies with audio and video based co-training. In: Pacific-Rim Conference on Multimedia (2009)
25. Lu, Z., Carreira-Perpinan, M.A.: Constrained spectral clustering through affinity propagation. In: IEEE Conference on Computer Vision and Pattern Recognition (2008)
26. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: International Joint Conferences on Artificial Intelligence (1981)
27. Myhill, D.: Bad boys and good girls? patterns of interaction and response in whole class teaching. *Br. Educ. Res. J.* **28**(3), 339–352 (2002)
28. Newman, M.E.J.: Modularity and community structure in networks. *Proc. Nat. Acad. Sci.* **103**(23), 8577–8582 (2006)
29. Pei, M., Dong, Z., Zhao, M.: Event recognition based on social roles in continuous video. In: IEEE International Conference on Multimedia and Expo (2013)
30. Qiu, J., Lin, Z., Tang, C., Qiao, S.: Discovering organizational structure in dynamic social network. In: IEEE International Conference on Data Mining (2009)
31. Ramanathan, V., Yao, B., Fei-Fei, L.: Social role discovery in human events. In: IEEE Conference on Computer Vision and Pattern Recognition (2013)
32. Rasheed, Z., Shah, M.: Movie genre classification by exploiting audio-visual features of pre-views. In: International Conference on Pattern Recognition (2002)
33. Ruhnau, B.: Eigenvector-centrality? a node-centrality. *Soc. Netw.* **22**(4), 357–365 (2000)
34. Shi, J., Tomasi, C.: Good features to track. In: IEEE Conference on Computer Vision and Pattern Recognition (1994)
35. Smola, A.J., Schölkopf, B.: A tutorial on support vector regression. *Stat. Comput.* **14**(3), 199–222 (2004)
36. Song, Y., Morency, L.-P., Davis, R.: Action recognition by hierarchical sequence summarization. In: IEEE Conference on Computer Vision and Pattern Recognition (2013)
37. Sugiyama, M.: Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis. *J. Mach. Learn. Res.* **8**, 1027–1061 (2007)
38. Wang, G., Gallagher, A., Luo, J., Forsyth, D.: Seeing people in social context: recognizing people and social relationships. In: European Conference on Computer Vision (2010)
39. Wasserman, S., Faust, K., Iacobucci, D.: *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge (1994)
40. Weng, C.-Y., Chu, W.-T., Wu, J.-L.: Rolenet: Movie analysis from the perspective of social networks. *IEEE Trans. Multimedia* **11**(2), 256–271 (2009)
41. Yanagawa, A., Chang, S.-F., Kennedy, L., Hsu, W.: Columbia university's baseline detectors for 374 Iscom semantic visual concepts. Technical report, Columbia University (2007)
42. Yang, T., Chi, Y., Zhu, S., Gong, Y., Jin, R.: A bayesian approach toward finding communities and their evolutions in dynamic social networks. In: SIAM Conference on Data Mining (2009)
43. Yilmaz, A., Shah, M.: Recognizing human actions in videos acquired by uncalibrated moving cameras. In: International Conference on Computer Vision iccv (2005)
44. Yilmaz, A., Shah, M.: A differential geometric approach to representing the human actions. *Comput. Vis. Image Underst.* **109**(3), 335–351 (2008)
45. Yu, T., Lim, S.-N., Patwardhan, K., Krahnstoeve, N.: Monitoring, recognizing and discovering social networks. In: IEEE Conference on Computer Vision and Pattern Recognition (2009)
46. Zhai, Y., Shah, M.: Video scene segmentation using markov chain monte carlo. *IEEE Trans. Multimedia* **8**(4), 686–697 (2006)
47. Zhang, D., Gatica-Perez, D., Bengio, S., McCowan, I.: Modeling individual and group actions in meetings with layered hmms. *IEEE Trans. Multimedia* **8**(3), 509–520 (2006)

Human-Centered Social Media Analytics

Fu, Y. (Ed.)

2014, VIII, 208 p. 97 illus., 51 illus. in color., Hardcover

ISBN: 978-3-319-05490-2