MINING REPRESENTATIVE ACTIONS FOR ACTOR IDENTIFICATION

Wenlong Xie, Hongxun Yao, Xiaoshuai Sun, Sicheng Zhao, Tingting Han, Cheng Pang

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
{wlxie, h.yao, xiaoshuaisun, zsc}@hit.edu.cn

ABSTRACT

Previous works on actor identification mainly focused on static features based on face identification and costume detection, without considering the abundant dynamic information contained in videos. In this paper, we propose a novel method to mine representative actions of each actor, and show the remarkable power of such actions for actor identification task. Videos are firstly divided into shots and represented by BoW based on spatial-temporal features. Then we integrate the prototype theory with SVM to rank the shots and obtain the representative actions. Our method for actor identification combines representative actions with actors' appearance. We validate the method on episodes of the TV series "The Big Bang Theory". The experimental results show that the representative actions are consistent with human judgements and can greatly improve the matching performance as complementary to existing handcrafted static features for actor identification.

***Index Terms*—** Learning Representative Actions, Actor Identification, Dynamic Information

1. INTRODUCTION

Video data contain rich information about movements of subjects and changes in the environment, which are highly important for many visual tasks, including action identification, indexing, retrieval of scenes and analysis of movies. When watching a specific movie, we can perceive various information of an actor, such as face, costume, shape, body language, speech rhythm, etc. Among them, we might be interested that what makes this actor impressive, does he/she have some representative actions? Thus the understanding of movie characters is a necessary and meaningful work.

Detecting and identifying actors is the key problem in movie analysis. Most existing actor identification methods mainly focused on static features, such as face identification and costume detection [1, 2, 3, 4, 5], which achieves high accuracy under ideal conditions. However, all these methods may fail when the appearances of actors change greatly over time [5]. In this paper, we try to mine some dynamic features contained in movies for actor identification. It is natural for us to take actions into account, which vary significantly depending on different actors and scenes. We try to answer two main questions: (1) Can we extract representative actions of each actor from a movie? (2) How can these actions help for actor identification?

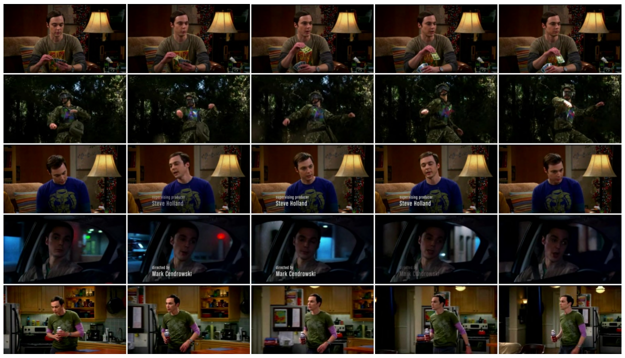

Fig. 1. What makes an actor impressive? These are some representative actions of Sheldon chosen manually. From top to bottom, the actions are playing cards, being shotted, shaking head, talking sideward, and walking, respectively.

2. RELATED WORKS

Existing actor identification approaches mainly used person specific static features, such as face detection. Everingham et al. [1] detected facial feature points to build a frontal faces descriptor, associate with scripts and subtitles to name and identify actors in TV videos. Building on the previous method, Sivic et al. [2] extended the coverage by using profile views. As their work significantly improved the accuracy of actor identification, they suggested future work focus on other nonfacial cues, such as hair and clothing. Indeed, Ramanan et al. [6] demonstrated that color histogram of body appearance can be used as a strong cue to group detected faces into tracks. They detected frontal faces, built a torso, face, and hair model for each detection, and then tracked actors using the body models. The experiment revealed that cues for nonfacial actor detector are available. Gandhi et al. [5] presented a generative appearance model to detect and name actors in movies, using maximally stable color regions (MSCR) [7]. By incorporating the costume of actors and representing the

This work was supported by the National Natural Science Foundation of China (No. 61472103) and Key Program (No. 61133003).

978-1-4799-9988-0/16/$31.00 ©2016 IEEE

heads and shoulders of actors as a constellation of optional color regions, their model was robust to changes in viewpoint and pose. However, when the appearances of actors change dramatically, the detection results are still unsatisfying. The authors suggested that future model might take the temporal coherence into account.

Our work in this paper also relates to some inspiring researches on person re-identification [8, 9] and action identification [10, 11]. Jain et al. [8] firstly used mid-level discriminative spatio-temporal patches to represent videos. Zhao et al. [9] learned mid-level filters from video patch clusters for person re-identification. Combined with existing handcrafted low-level features, their method significantly improved the performance of person re-identification. Laptev [10] proposed a new interest point detector which can find local image features in space-time characterized by a high variation of the image values in space and non-constant motion over time. The resulting points correspond to interesting events in video data. Laptev [11] then used the spatio-temporal interest points for action recognition in realistic human actions. Their method for action classification extends successful image recognition methods to the spatio-temporal domain and achieves best up to date recognition performance on a standard benchmark.

Motivated by previous works, we also try to utilize other new level features for person identification. Our work begins with the question "what makes an actor impressive in a movie" and purpose on mining dynamic information for actor identification, which is closely related to work in [5]. While Grandhi suggested to design models coherent with temporal changes, we try to obtain the temporal changes and get use of them. We invited a group of students to take a test, giving each of them video clips of TV series "The Big Bang Theory". The participants in this test only needed to mark if a video clip contains an actor's representative actions or not. After the test, we got each actor a group of specific actions, which people believe can make the actor impressive and distinguishable from others. We name these actions an actor's representative actions. Some results are shown in Fig. 1.

Different from action identification in [12], our representative action mining method focuses on finding similar actions in video clips and ranking them by representativeness. It is difficult to define and calculate an action's representativeness. To resolve this issue, we propose a representative action mining method by integrated the prototype theory with the classic support vector machine (SVM).

Prototype Theory The prototype theory [13, 14] in cognitive science, states that categories tend to be defined in terms of prototypes or prototypical instances that contain the attributes most representative of items inside and least representative of items outside a category. Sun et al. [15] introduced prototype in their visual representativeness model and got promising results. Extended from above work, our approach could obtain representative actions which are consistent with human judgements.

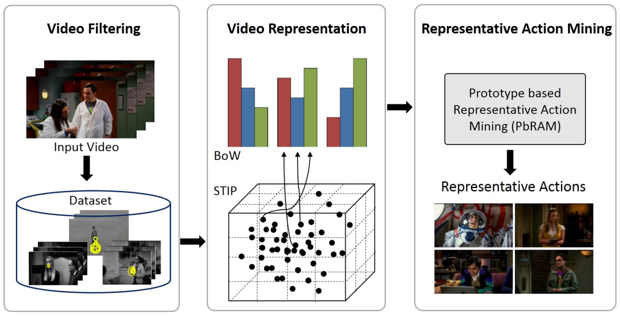

Fig. 2. The framework of the proposed method for mining representative actions

3. MINING REPRESENTATIVE ACTIONS

Given a movie or episodes of TV series, firstly we divide them into video clips and select the ones that only contain a specified actor. Then we represent these clips by BoW, and use spatio-temporal interest points (STIP) [11] for spatio-temporal feature extraction. Finally, we propose a novel method, named prototype based representative action mining (PbRAM), to mine representative actions for each actor. The framework of the proposed method is shown in Fig. 2.

3.1. Video Preprocessing

In this paper, we choose TV series "The Big Bang Theory" Season 5 and Season 7 for mining representative actions of actors. In order to get the actions of each actor, videos are automatically divided into clips by shot. Totally we get about 16800 clips from the TV series. For each actor, we need to select the clips that contain them. The video clips are preprocessed by the following steps.

Table 1. Numbers of actions (video clips) in our database

Actor	Sheldon	Leonard	Penny	Raj	Howard
#Video	1936	1019	872	688	652

3.1.1. Video Filtering by Face Detection

As face detection on all video frames is time-consuming, we extract key frames of each clip and detect faces in them to filter irrelevant videos. Then for each actor, we can obtain a number of video clips containing their own actions. Table. 1 shows the actions numbers (video clips) of each actor, where #Video means number of actions. It is interesting to find that these numbers somehow reveal the contribution of each actor to this TV drama.

3.1.2. STIP based Video Representation

We detect spatio-temporal interest points (STIP) and compute corresponding local spatio-temporal descriptors for all

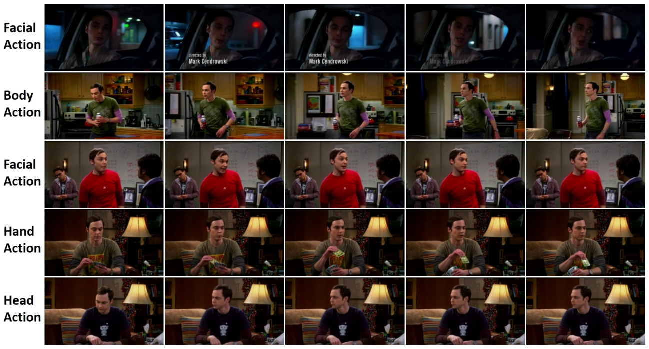

Fig. 3. Representative actions of actors mined by our method. From first row to last row: talking to sideward, walking, saying words “Bite me?” which is kind of provocation, playing cards, looking around.

video clips to represent actions. Following the work of Laptev [11], we detect the spatial-temporal extremum points from the video frames, and then extract Histograms of Oriented Gradients (HOG) feature and Histograms of Optical Flow (HOF) feature for the descriptors. Here we also filter the video clips that have too few or too many changes in time, in which the actors have too subtle or too complex action movements. Using BoW, we turn descriptors of a same clip into a 1000 dimensional histogram.

3.2. Representative Action Mining Method

Given the BoW representations of an actor’s actions, our PbRAM method cluster them into centers using k-means with Cosine distance as the clustering metric. These centers are set as the initial prototypes of this actor’s actions. We then train a SVM classifier for this actor, using the prototypes of this actor as positive exemplars and the prototypes of other actors as negative exemplars. The SVM classifier will mark the actions, which are the most or the least similar to the centers of action prototypes, the maximum positive or the minimum scores respectively among all action scores. Obviously, it is consistent with the definition of prototype theory. Algorithm. 1 shows the representative action mining method integrated with the prototype theory and the classic SVM method. In the experiment we set $N_a = 5, k = 5, N_r = 10, \theta = 60\%$, which means we need to mine the top 10 representative actions for the 5 actors in movie. The mined representative actions of Sheldon are shown in Fig. 3. Compared with students test result in Fig. 1, representative actions mined by our method are consistent with human judgements on representativeness. Sheldon is one of the main actors in “The Big Bang Theory”, he always has some little nervous movements no matter when he is talking or walking, and this point impresses us. For more details about the mined representative actions, please check the supplementary materials.

Algorithm 1: Prototype based representative action mining

```
Input: Actor number $N_a$, top $N_r$ representative actions need to
       mine, BoW representation of each actor's actions
       $H_i = \{hist_j\}_{j=1}^{N_i}, i = 1, ..., N_a$, number of clustering
       centers $k$, threshold $\theta$ to split prototypes with other actions
Output: Representative actions index matrix $\mathbf{R}(N_a, N_r)$
for $i = 1, ...N_a$ do
    Run $k$-means on $H_i$ to obtain cluster centers $C_i = \{c_j\}_{j=1}^{k}$;
end
Unite all the prototypes $C = \{C_i\}_{i=1}^{N_a}$;
for $i = 1, ...N_a$ do
    $Classifier_i \leftarrow$ trainSVM$(C_i, C - \{C_i\}, cosine)$;
    $score_i \leftarrow$ test all item in $H_i$ with $Classifier_i$;
    Descending sort $score_i$;
    $N = \max\{j\}$, s.t. $score_i(j) > 0$;
    if $N \geqslant N_i * \theta$ then
        $\mathbf{R}(i, N_r) \leftarrow$ index of top $N_r$ ranking actions of $score_i$;
        break;
    end
    $C_i \leftarrow$ corresponding top $k$ score actions in $H_i$;
end
return $\mathbf{R}(N_a, N_r)$;
```

4. ACTOR IDENTIFICATION WITH REPRESENTATIVE ACTIONS

In this section, we introduce how to use the mined representative actions for actor identification. The appearance based identification method of Sivic [2] is implemented as baseline. We then utilize our resulting representative actions as complementary to the method. Given prototypes of an actor’s actions or appearance, the similarity between test item $\mathbf{x}^*$ and the underlying prototype $\mathbf{r}$ is:

$$score_i(\mathbf{x}^*, \mathbf{r}) = \exp\{-\lambda D(\mathbf{x}^*, \mathbf{r})\}, \quad (1)$$

where λ is a scaling constant to keep $score_i(\mathbf{x}^*, \mathbf{r})$ in a reasonable interval. $D(\mathbf{x}^*, \mathbf{r})$ is the distance between the prototype and the test items. We use Euclidean distance for appearance matching, and Cosine distance for representative action matching.

According to Eq.(1) we get two scores, one is the appearance identification result, and the other is the representative action matching result. The final score of actor identification is a weighted sum of these two scores:

$$score(\mathbf{x}^*) = \omega_1 * score_1(\mathbf{x}^*, \mathbf{r}) + \omega_2 * score_2(\mathbf{x}^*, \mathbf{r}), \quad (2)$$

where $0 \leqslant \omega_1, \omega_2 \leqslant 1$ and $\omega_1 + \omega_2 = 1$.

5. EXPERIMENTS

To evaluate the effectiveness of the proposed method, we conduct experiments on TV series “The Big Bang Theory”.

5.1. Dataset

To our knowledge, there is no public dataset available for mining representative actions of actors. Due to this reason, for a

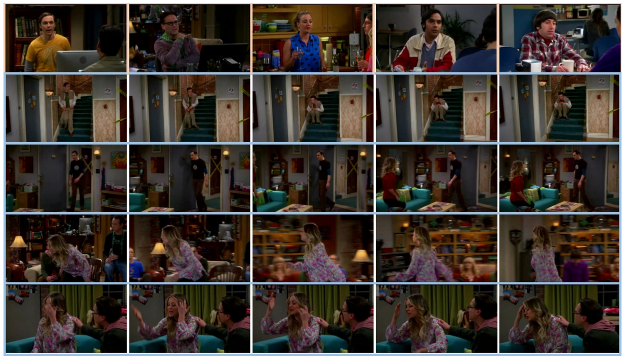

Fig. 4. Actor identification examples of our method. The first row shows examples of actors that can be identified by face appearance or representative action matching. Each of the second to the last row shows an example that an actor can only be identified by representative action matching.

detailed evaluation of our method. We set up a dataset, named **BigBangActions**, which is composed of video clips from TV series “The Big Bang Theory”. Each video clip corresponding to an action is manually labelled with the actor name and consists of 30∼300 frames. There are many actors in TV series ” The Big Bang Theory ” in fact. A significant factor we need consider is that an actor’s actions should be various and abundant for mining representative actions, which leads to only five main actors remained in the forthcoming BigBangActions dateset. Each actor in the dataset has around 1000 selected actions.

5.2. Results

We ran our representative actions mining method on randomly selected 50% videos in the dataset, and identified actors in the other 50% videos. Our method performs best when $\omega_1 : \omega_2 = 2 : 3$, achieving an accuracy of 75.32%. The identification results on all five actors are presented in Fig. 5, which are consistent on different actors.

Examples of the successful identification results are shown in Fig. 4, which demonstrate our method performs well even with severe temporal changes. In the first row of Fig. 4, all the actors show their frontal faces and the shot is relatively stable in temporal order. In the second row, Sheldon is far away from camera, which causes face detection on him fail. In the third row, Sheldon walks into Penny’s house quickly and the camera even could not obtain his face. However, Sheldon’s actions in the above two video clips match his representative actions, which help us to successfully identify him. The fourth row and last row also show examples that Penny’s faces are hard to been detected in the situation when her appearances change greatly over time or in the presence of occlusion. As Penny is straightforward and impatient, her actions are usually finished in a wide range. It is fortunate that actions in the last two rows of Fig. 4 are remarkable, and could be easily matched to Penny’s representative actions. Video clips of Fig. 4 show specific information about how the actor appearance changes, please check the supplementary materials.

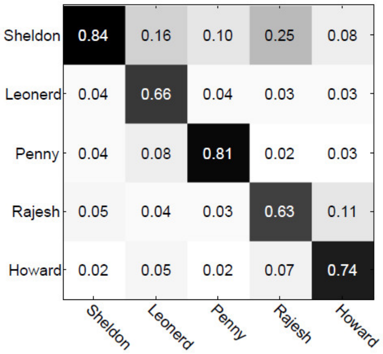

Fig. 5. Confusion matrix of our proposed method on the five actors for actor identification.

The performance comparison between our method and the baseline based actors’ appearance [2] is shown in Table 2. As we can see, our method greatly improves the average accuracy of actor identification, which demonstrates that representative action is promising and could greatly help for actor identification as a complementary dynamic feature for static features.

Table 2. Performance comparison between the proposed method and the baseline based on actors’ appearance [2].

Method	Accuracy(%)
Sivic et al. [2]	59.78
Representative Actions	34.81
Combination	**75.32**

6. CONCLUSION

In this paper, we tried to discover dynamic features for actor identification from video data. To accomplish this work, a new dataset, named BigBangActions, for mining representative actions of actors was constructed and will be released soon. We introduced the concept of representative actions and proposed a prototype based representative action mining (P-bRAM) method to mine them from videos. The representative actions mined by our method are consistent with human judgments. Using the representative actions as complementary to appearance detection, we improved the performance of actor identification, especially in situations when the appearance of an actor changes quickly. The experimental results demonstrated the effectiveness of the proposed method. In the future work, we would like to improve the mining method and explore how we can make better use of representative actions to acquire more comprehensive understandings of both static and dynamic semantics in story videos. One of our ongoing work aims at utilizing representative actions in specific person retrieval [16] and personality analysis.

7. REFERENCES

[1] M. Everingham, J. Sivic, and A Zisserman, ""Hello! My name is... Buffy" – automatic naming of characters in TV video," in *British Machine Vision Conference*, 2006.

[2] Josef Sivic, Mark Everingham, and Andrew Zisserman, "who are you?-learning person specific classifiers from video," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1145–1152.

[3] Piotr Bojanowski, Francis Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid, and Josef Sivic, "Finding actors and actions in movies," in *IEEE International Conference on Computer Vision*, 2013, pp. 2280–2287.

[4] Enrique G Ortiz, Alan Wright, and Mubarak Shah, "Face recognition in movie trailers via mean sequence sparse representation-based classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3531–3538.

[5] Vineet Gandhi and Remi Ronfard, "Detecting and naming actors in movies using generative appearance models," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3706–3713.

[6] Deva Ramanan, Simon Baker, and Sham Kakade, "Leveraging archival video for building face datasets," in *IEEE International Conference on Computer Vision*, 2007, pp. 1–8.

[7] P-E Forssén, "Maximally stable colour regions for recognition and matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[8] Arpit Jain, Abhinav Gupta, Mikel Rodriguez, and Larry S Davis, "Representing videos using mid-level discriminative patches," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2571–2578.

[9] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Learning mid-level filters for person re-identfiation," in *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, USA, June 2014.

[10] Ivan Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[11] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld, "Learning realistic human actions from movies," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[12] Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International journal of computer vision*, vol. 79, no. 3, pp. 299–318, 2008.

[13] Eleanor Rosch, "Cognitive representations of semantic categories," *Journal of experimental psychology: General*, vol. 104, no. 3, pp. 192, 1975.

[14] Eleanor Rosch, "Principles of categorization," *Concepts: core readings*, pp. 189–206, 1999.

[15] Xiaoshuai Sun, Xin-Jing Wang, Hongxun Yao, and Lei Zhang, "Exploring implicit image statistics for visual representativeness modeling," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 516–523.

[16] Sicheng Zhao, Lujun Chen, Hongxun Yao, Yanhao Zhang, and Xiaoshuai Sun, "Strategy for dynamic 3d depth data matching towards robust action retrieval," *Neurocomputing*, vol. 151, pp. 533–543, 2015.