Person Identity Clustering in TV Show Videos

Yina Han*†, Guizhong Liu*

*School of Electrical and Information Engineering, Xi'an Jiaotong University, Xi'an, P.R.China
Email:ynhan@mailst.xjtu.edu.cn, liugz@mail.xjtu.edu.cn

Gerard Chollet†, Joseph Razik†

†CNRS-LTCI, TELECOM-ParisTech
Paris, France
Email: {han, chollet, razik}@enst.fr

Keywords: Identity clustering, shot layer matching, face detection, mean shift, significant colour

Abstract

The goal of this work is to cluster all occurrences of TV show characters into different sets with respect to their own identity. A two stage method is presented in this paper. First shot layer processing is performed to gather close-ups with the same central person together and to differentiate far full views from close-up views, hence complicated inter shot facial feature matching is avoided. Then their exact position is localized by face detection and tracking for close-up views. For far full views significant clothing colour recovering and matching is presented. Experiments are presented on a segment from the famous French TV show 'Le Grand Echiquier' with satisfactory results.

1 Introduction

The goal of this work is to cluster all occurrences of individuals into different sets with respect to their own identities for TV show videos. By identity clustering we mean to gather the same person appeared at different time with different views together and bound their exact position in the image. There are many applications of such a capability, for example: provide a set of exemplars with varied poses, scales and expressions for specific person recognition, retrieval [6] and behaviour analysis.

One approach to this problem is to use face detection [9] and face matching [1, 4, and 6]. However, faces may go undetected if the person turns towards profile and back to frontal [6], besides even for the detected faces, matching is notoriously difficult for the variation in pose, partial occlusion and expression can exceed that due to identity [6].

Hence in [1, 4, 6] sets of faces for each person are first gathered in shots using tracking. And then SIFT based local facial feature matching is adopted for inter shot matching. However, as it is said in [4], errors may be introduced by incorrect localization of the features and SIFT descriptor at several scales is a great computation cost. Moreover, for the far view shots, no face can be detected due to the resolution decline of the whole image.

In [7], they went beyond face detection and re-detected people where their face detection is missing by explore other cues such as clothing colour. Each person is modelled as a pictorial structure [5] with three rectangular parts corresponding to hair, face and clothing region with a single colour appearance model for each part. But the restriction of stable background, unchanged scale and view point can not be met in our situation.

In this paper, by fully exploring the production styles of TV show video, we propose a two step method for person identity clustering. As shown in Figure 1, first shot layer processing is performed to gather close-ups with the same central person together and to differentiate far full views from close-up views. Then object layer processing, including face detection and tracking for close-up view, and significant clothing colour matching for far full view, is followed to further localize their exact position.

Compared with previous work [4, 6, and 7], we bring two areas of novelty: First, for close-up view, we introduce shot layer matching to firstly gather the shots that contain the same central person together, and accordingly avoiding complicated inter shot facial feature matching [1, 4, and 6]; Second, an unsupervised significant colour recover and matching strategy is proposed, which can solve the challenge problem beyond [7] that is variation in scale, view point and not unique clothing colour.

The rest of this paper is organized as follows: Section 2 discusses shot layer processing, Section 3 describes object layer processing for close-up view. Significant colour mode construction and matching strategy for far full view is introduced in Section 4. Experimental and some conclusions are presented in Section 5 and Section 6 respectively.

2 Shot layer processing

Unlike movie stories, TV show program exhibits limited and compact shot types since the whole process takes place in a fixed studio around a guest and alternates interviews and informal discussions among the guest principal, the conductor and other invited people. Correspondingly, from the aspect of vision, the whole video sequence is composed of alternate close-up views of several specific persons and far full views of the whole studio. In order to fully explore this character shot layer processing is performed first.

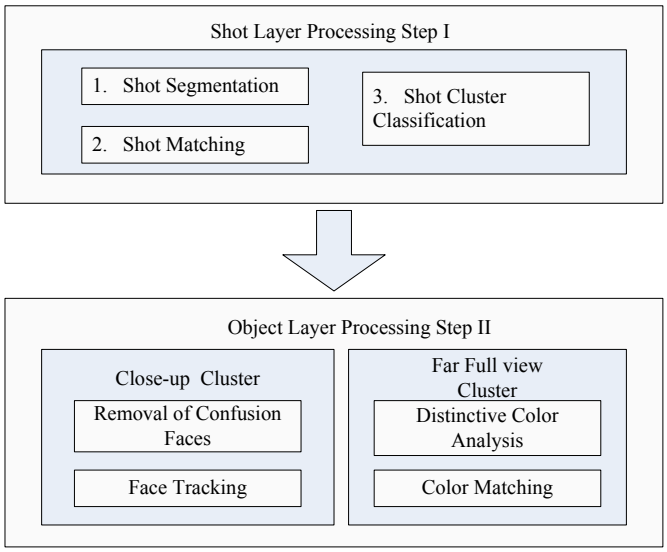


Figure 1: Framework for person identity clustering.

2.1 Shot segmentation

Shot detection is fairly mature and a number of reliable shot detection algorithms have been built. In our case a simple method based on RGB colour histogram difference between consecutive frames is adopted.

For simplicity, given a frame defined by RGB colour axes, the colour histogram is obtained by discretizing the R, G, B colour components respectively and counting the number of times each discrete component occurs in the image array. And for each frame, its colour histogram H_i is compared with the previous one H_{i-1} by colour histogram intersection (CHI) [7]:

$$CHI_i = \sum_{r,g,b} \min\{H_i(r,g,b), H_{i-1}(r,g,b)\} \quad (1)$$

CHI_i takes value between 0 and 1, being 1 when $H_i = H_{i-1}$.

An example of colour histogram intersection for a sequence of 4965 frames with the spotted ground truth by red stars is shown in Figure 2. As we can see the shot boundary is notable.

2.2 Shot matching

Similar to consecutive frames comparison, each shot is represented by the average of the frame histograms belonging to it. And the similarity between two shot histograms (SH) is measured by their histogram intersection:

$$SCHI_{i,j} = \sum_{r,g,b} \min\{SH_i(r,g,b), SH_j(r,g,b)\}, \ 1 \le i, j \le NS \quad (2)$$

where NS is the number of shots. As Figure 3 shown, the measure of intersection for shots of the same kind, which are spotted by red stars, is notable and robust.

2.3 Shot cluster classification

From the perspective of view type, shot layer matching provides two main clusters: clusters with close-up view for specific persons, as shown in Figure 3 (a)-(e), and cluster of far full view of the whole studio as shown in Figure 3 (f). We differentiate the two by face detection results. For each cluster, OpenCV implementation of the method of Viola and Jones [9] for frontal face detection is run on every frame, and due to the poor visual resolution, the far full view cluster is differentiated as the one with almost no detected faces.

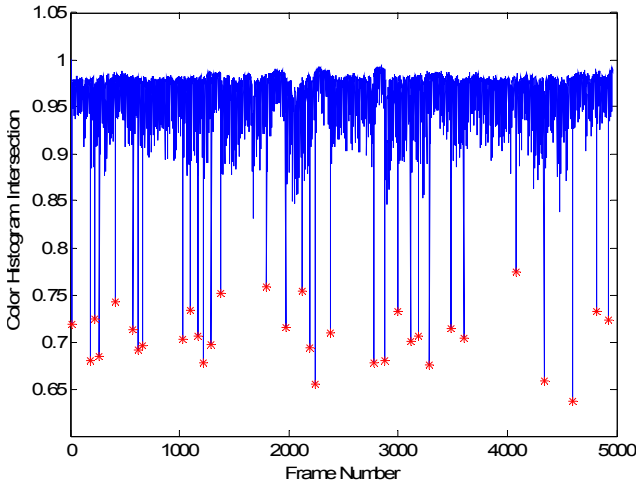


Figure 2: Colour histogram intersection for consecutive frames. Red stars correspond to the shot boundaries.

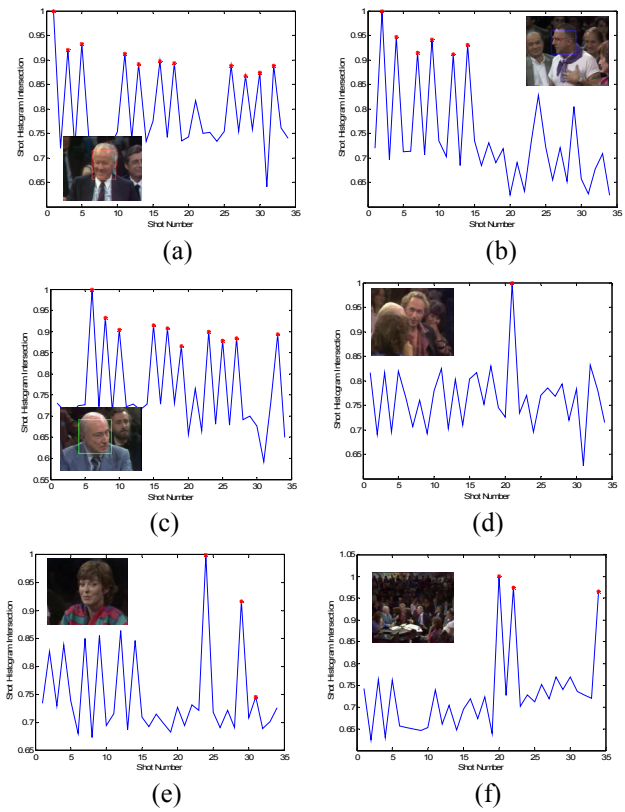


Figure 3: Shot similarity measured by histogram intersection for six shots. Red stars correspond to shots of the same kind. Thumbnails show representative frame of shot to be matched.

3 Object layer processing for close-up clusters

The goal of this section is to localize the exact face region of central persons appeared in close-up views. This can not be solved directly by face detection, because on one hand face detection may introduce noisy faces from background people who also face the camera as shown in Figure 4(a), on the other hand there may be drop outs as the person turns towards profile and back to frontal [6] as shown in Figure 4(b). Hence faces cleansing and tracking is proposed.

(a) (b)

Figure 4: Two main problems caused by single face detection.

3.1 Noisy faces cleansing

The noisy faces are coarsely cleansed by the prior knowledge of the scale and position that a central person should appear with in close-up view, but with a more strict low false positive principle to ensure all the noisy faces can be removed. Once faces for a central person are determined, the missing ones can be redetected by tracking (both forward and backward).

3.2 Intra cluster person tracking

Here mean-shift colour tracker is adopted [3]:

Let $\{x_i^*\}_{i=1\ldots n}$ be the pixel locations in the detected face region. The function $b: R^2 \to \{1\ldots m\}$ associates to the pixel at location x_i^* the index $b(x_i^*)$ of its bin in the quantized colour space. The target model $\{\hat{q}_u\}_{u=1\ldots m}$ cantered at $\hat{y}_0$ in previous frame is then computed as

$$\hat{q}_u = C\sum_{i=1}^{n} k\left(\left\|y_0 - x_i^*\right\|^2\right)\delta\left[b(x_i^*) - u\right] \tag{3}$$

The complete target localization algorithm is as below:

1. Initialize the location of the target in the current frame with $\hat{y}_0$, compute $\{\hat{p}_u\}_{u=1\ldots m}$ with the same kernel profile $k(x)$

$$\hat{p}_u(y) = C_h\sum_{i=1}^{n_h} k\left(\left\|y - x_i\right\|^2\right)\delta\left[b(x_i) - u\right] \tag{4}$$

2. Derive the weights $\{w_i\}_{i=1\ldots n_h}$ according to

$$w_i = \sum_{u=1}^{m}\sqrt{\frac{\hat{q}_u}{\hat{p}_u(\hat{y}_0)}}\delta\left[b(x_i) - u\right] \tag{5}$$

3. Find the next location of the target candidate according to

$$\hat{y}_1 = \frac{\sum_{i=1}^{n_h} x_i w_i g\left(\left\|\frac{\hat{y}_0 - x_i}{h}\right\|^2\right)}{\sum_{i=1}^{n_h} w_i g\left(\left\|\frac{\hat{y}_0 - x_i}{h}\right\|^2\right)} \tag{6}$$

4. If $\|\hat{y}_1 - \hat{y}_0\| < \varepsilon$ stop.

Otherwise, set $\hat{y}_0 \leftarrow \hat{y}_1$ and go to step 1.

Here, kernels with Epanechnikov profile

$$k(x) = \begin{cases} \frac{1}{2}c_d^{-1}(d+2)(1-x) & if\ x \le 1 \\ 0 & otherwise \end{cases} \tag{7}$$

is adopted. In this case, the derivative of the profile, $g(x)$ is constant and (6) reduces to

$$\hat{y}_1 = \frac{\sum_{i=1}^{n_h} x_i w_i}{\sum_{i=1}^{n_h} w_i} \tag{8}$$

i.e., a simple weighted average.

4 Significant colour matching for far full view

Given the bounded faces for each close-up cluster, the goal here is to localize its corresponding far full view appearance, as shown in Figure 5. Due to face detection is completely missing in far full views, significant clothing colour cues are explored. Significant colours correspond to high density regions in this space [2], which have a better visual invariance to changes of view points and scales (namely from close-up view to far full view as shown in Figure 5). Hence first the centres of the high density colour for the clothing regions are recovered from close-up views by mean shift procedure. And then people of far full view is located and identified to their corresponding close-ups by a flexible colour matching strategy.

Figure 5: Correspondence between close-up and far full view.

4.1 Significant clothing colour recovery from close-up view

For each representative frame $F_i, 1 \le i \le M$ which is randomly chosen from the close-up views of a specific person, and M is the number of persons to be matched, his/her clothing region $CR_i, 1 \le i \le M$ can be predicted by its relative location and scale to the face region [4, 6, 7]. After mean shift procedure, which is a nonparametric clustering technique and neither require prior knowledge of the number of clusters, nor constrain the shape of the clusters [2], the clothing region is quantized into several centres $\{C_i^1, \ldots C_i^{N_i}\}, 1 \le i \le M$ with their

corresponding proportions $\{p_i^1,\ldots p_i^{N_i}\}, 1 \le i \le M$ in RGB space, where N_i is the number of colours for the i -th region CR_i .

4.2 Colour matching

Given the clothing colour centres and their corresponding proportion, a set of representative significant colours are chosen for each person. Generally speaking, colours with larger proportion are more stable while small proportion ones may either come from background noise or become invisible as the scale and view point changes. Taking the set of significant colours for all the persons to be matched as centres, and each pixel in a far full view frame is classified into one of them according to the measure of Euclidean distance in RGB colour space with the nearest neighbour principle. And the result colour distance images for each person provide a flexible way to choose the colours not only significant in proportion but also in distinguish ability.

5 Experimental results

The proposed method has been applied on a segment from the famous French TV show 'Le Grand Echiquier'. The entire segment has 4965 frames and 34 shots arising from a far full view as shown in Figure 6(f), and five close-up views as shown in Figure 6(a)-(e).

Shot detection: The ROC curve for shot detection is shown in Figure 6, from which we can see that shots can be well segmented with 100% precision rate and 0% false positive rate by selecting proper threshold, and 0.8 is chosen here.

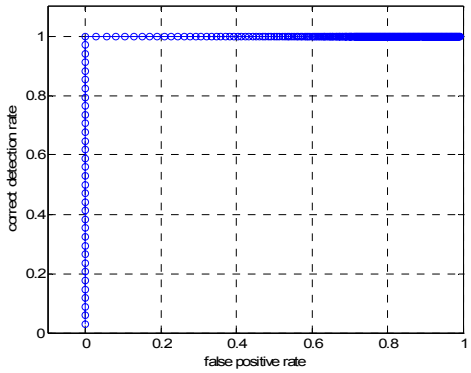


Figure 6: ROC curves for shot detection.

Shot matching: Given a shot, the matching results are evaluated by the precision/recall curve as shown in Figure 7. Still the characteristic of these curves indicates that proper threshold can give well matching result with 100% precision rate and 100% recall rate, which guarantees well identity clustering for close-up view.

Object layer localization: The object layer localization is performed for three frequent appeared people. As shown in Figure 7, they are Gerard Oury (Figure 7(a)), Coluche (Figure 7(b)) and Jacques Chancel (Figure 7(c)).

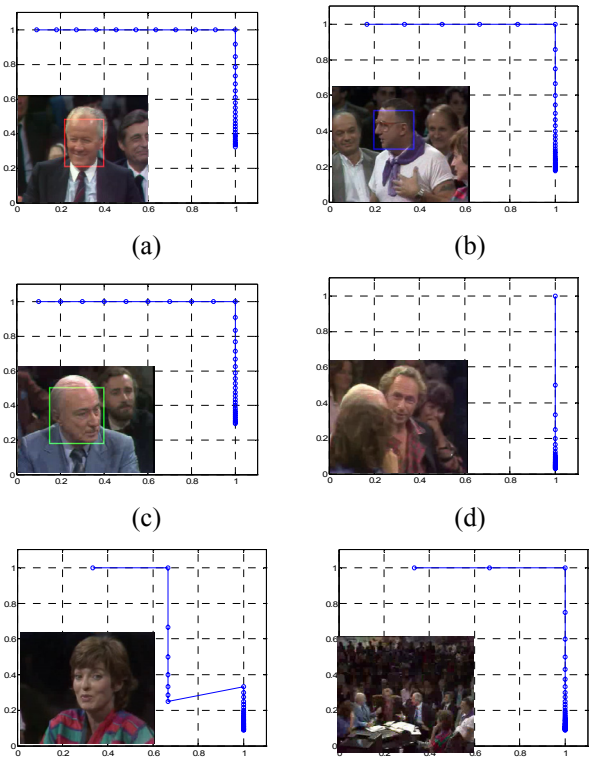


Figure 7: View layer retrieval results. The graphs show precision (y-axis) vs. recall (x-axis). Thumbnails show representative frame of shot to be matched.

	MB	FD
Jacques	82.9%	100%
Gerard	100%	97.7%
Coluche	85.9%	100%

Table 1: Precision rate for close-up view face localization.

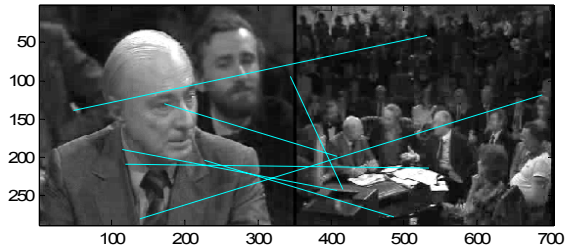
Close-up view: since we use tracking to localize faces, the position of the reference region is crucial. Localization results by tracking from the manually bounded (MB) first frame, and tracking combined with the cleansed face detection (FD) results are compared in Table 1.

Far full view: SIFT based matching are first conducted. As shown in Figure 7, due to the decline of resolution in far full view frame, this matching is not reliable.

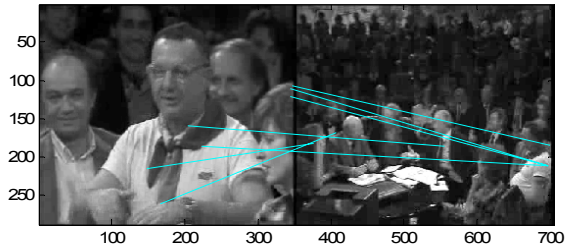
Our proposed colour matching are implemented as: First colours are recovered from the lower 1/3 parts of the close-up views, which is mainly occupied by clothing, by mean shift cluster with Epanechnikov profile as shown in Equation (7) and 0.1 cluster band width. Then for each person, we choose the first two colour centres with the largest proportions since they cover the majority of the region, as shown in Table 2. Finally, matching is calculated for the lower half part of the far full view as shown in Figure 8(a). For each person in the far full frame two distance images corresponding to his two

significant colours are achieved. After median filtering as shown in Figure 8(b-d), not both of the results can provide meaningful differentiation, that's because one significant colour recovered from close-up view may become invisible in far full view due to the change of view point as Figure 8(c) shown; or the colour may be not unique in the scene as Figure 8(d). Hence for the two distance images, our localization is carried out on the one with a single compact and notable region. And this region is bounded as the objective as Figure 8(b)-(d) shown.

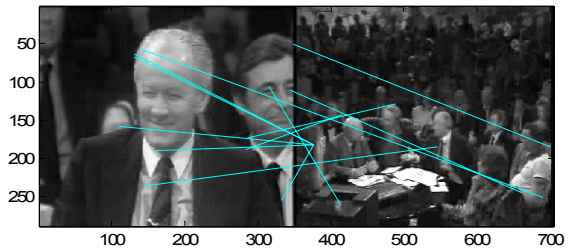
Once the corresponding colour objective with appropriate horizontal and vertical expansion is localized for the first frame as shown in Figure 10, the same mean shift tracking procedure is carried out for the rest frames with 100% precision.



(a)



(b)



(c)

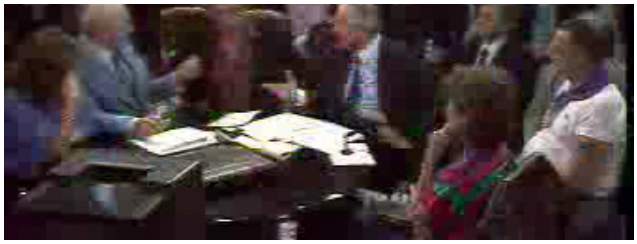
Figure 8: SIFT matching between close-up view and far full view

6 Conclusions

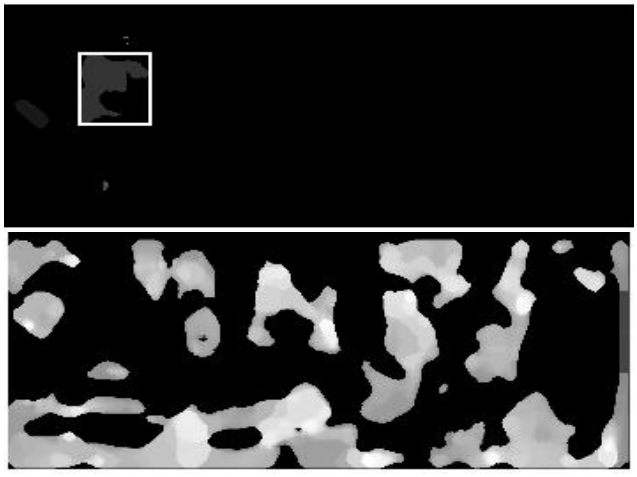
In this paper, we propose a two stage method to cluster and localize TV show characters with varied pose, expression and scale into different sets with respect to their own identity. Compared to traditional methods, where person clustering is mainly implemented by face detection and matching, we fully explore the shot characteristic of videos, shot layer matching and tracking is proposed to differentiate and localize person. Shot related methods are relative mature and simple, hence satisfactory result can be achieved. For the shot where no face can be detected, mean shift clustering based colour model is proposed. The experiments show that our model is more flexible to handle various situations. And future work is to use the face exemplars with varied poses, scales and expressions provided by this work for further person recognition and facial feature analysis.

	Colour I	Proportion	Colour II	Proportion
Jacques	0.1499 0.1386 0.1757	76.3%	0.5583 0.5983 0.6587	10.7%
Gerard	0.2819 0.326 0.423	84.9%	0.0946 0.1106 0.1369	12.4%
Coluche	0.344 0.315 0.3017	41.2%	0.795 0.792 0.814	41.2%

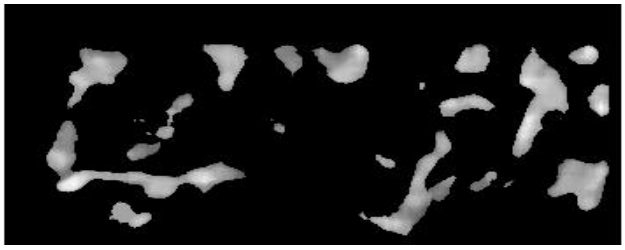
Table 2: Significant colours and their corresponding proportions

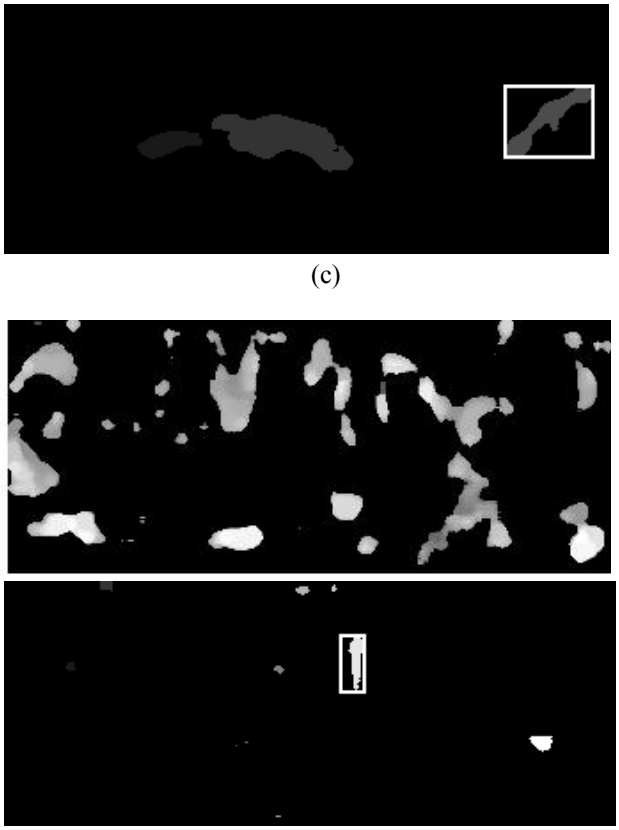


(a)



(b)





(c)

(d)

Figure 9: (a) is the objective region. (b) - (d) is the negative logarithm colour distance for Gerard, Claude and Jacques. The brightness indicates the similarity to that colour.

Figure 10: Localization result for a far full view frame.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable and constructive comments that helped improve the presentation of this paper substantially. This work is supported in part by National 973 Project No.2007CB311002，National Natural Science Foundation of China (CNSF) Project No.60572045, the Ministry of Education of China Ph.D. Program Foundation Project No.20050698033, China Scholarship Council, the K-space project of Europe, and Infom@gic project of France.

References

[1] L. Ballan, M. Bertini, A. Del Bimbo, W. Nunziati. “Soccer players identification based on visual local features”, *ACM international conference on image and video retrieval*, pp. 258-265, (2007).

[2] D. Comaniciu, P. Meer. “Mean shift: a robust approach toward feature space analysis”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24(5)**, pp. 603-619, (2002).

[3] D. Comaniciu, V. Ramesh, P. Meer. “Kernel-based object tracking”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25(5)**, pp. 564-577, (2003).

[4] M. Everingham, J. Sivic, and A. Zisserman. “”Hello! My name is…Buffy”- Automatic naming of characters in TV video”, *Proceedings of British Machine Vision Conference (BMVC)*, (2006).

[5] P. Felzenszwalb, D. Huttenlocher. “Pictorial structures for object recognition”, *International Journal of Computer Vision*, **61(1)**, pp. 55-79, (2005).

[6] J. Sivic, M. Everingham, and A. Zisserman. “Person spotting: video shot retrieval for face sets”, *International Conference on Image and Video Retrieval (CIVR)*, pp. 226-236, (2005).

[7] J. Sivic, C. L. Zitnick, and R. Szeliski. “Finding people in repeated shots of the same scene”, *Proceedings of British Machine Vision Conference (BMVC)*, (2006).

[8] M. J. Swain, D. H. Ballard. “Colour indexing”, *International Journal of Computer Vision*, **7(1)**, pp. 11-32, (1991).

[9] P. Viola, M. Jones. “Robust real-time fce detection”, *International Journal of Computer Vision*, **52(2)**, pp. 137-154, (2004).