Comparison of linear feature spaces for classification of face sequences in movie videos

Takuya Kobayashi† and Akinori Hidaka † and Takio Kurita ‡

†:University of Tsukuba, taku-kobayashi,hidaka.akinori@aist.go.jp
‡:Institute of Advanced Industrial Science and Technology (AIST),
Neuroscience Research Institute, takio-kurita@aist.go.jp

Abstract We consider classification problem of face sequences extracted from actual movie videos. At first all faces are extracted from each frame of the given movie videos by applying the popular face detector proposed by Viola and Jones. Then they are merged as a face sequences if the faces in the consecutive frames belong to the same shot and have similar size and location. Histogram of Oriented Gradients (HOG) features are extracted from each face image in the sequences and they are used to compare the similarity of the face sequences. In this paper, we compare the performance of the several dimension reduction methods, such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Locality Preserving Projection (LPP).

1 Introduction

In this paper, we consider classification problem of face sequences extracted from actual movie videos. The classification of persons in video such as movies, dramas, and news programs is important for understanding video contents. It enables many multimedia applications, for example video indexing, automated face annotation. However face sequence classification is challenging because appearance of faces in video has a lots of variations such as illumination, direction, expression.

At first, all faces are extracted from each frame of the given movie videos by applying the popular face detector proposed by Viola and Jones[1]. Then they are merged as a face sequences if the faces in the consecutive frames belong to the same shot and have similar size and location. We extracted Histogram of Oriented Gradients (HOG)[2] features from each face images.

To evaluate the similarity of the face sequences, it is important to construct a proper features space in which the features in the face sequences of the same person becomes close and the features of the different persons becomes apart. In this paper, we compare the performance of the several dimensionality reduction methods, such as Principal Component Analysis (PCA), Locality Preserving Projection (LPP)[3], and Linear Discriminant Analysis (LDA). PCA is one of the well known techniques for dimensionality reduction. In this paper, three different feature spaces are constructed by using PCA. The first feature space is constructed by applying PCA to a subset of all face images in all face sequences. This feature space reflects the variations of face appearances. We call this feature space PCA-ALL. After this we regarded one face sequence as same class samples. The second one is obtained by replacing the total covariance matrix of PCA-ALL with the within class covariance matrix. The constructed feature space reflects the variations in the face sequences. We call this feature space PCA-WITHIN. The last one is obtained by replacing the total covariance matrix of PCA-ALL with the between class covariance matrix. The constructed feature space reflects the variations of the average faces of the face sequences. We call this feature space PCA-BETWEEN. LPP is similar to PCA and projection matrix is calculated by Laplacian matrix of the data. This projection optimally preserves local neighborhood information. LDA is also applied to the face sequences by assuming different sequences are different classes.

The similarity between a pair of face sequences is evaluated by using Euclidian distance and discriminant criterion. Euclidian distance is the simplest measure. Discriminant criterion is defined as the ratio of the variance and the between-class variance and indicates the degree of separation of the two set of feature vectors. Thus we can decide two sequences are obtained from the same person if the similarity is sufficiently high.

2 Previous Works

Face information is important in videos. By extracting face sequences from videos, we can realize multimedia applications such as face retrieval, face annotation, video authoring, etc. For automatic casting, groups of similar faces are represented by key faces such as principal actors. We have to efficiently form clusters of face sequences.

Foucher *et al.* proposed a video indexing system that aims at indexing large video files in relation to the presence of similar faces [4]. The near-frontal view faces are detected by a cascade of weak classifier and tracked through a particle filter. For each trajectory, a representative sample composed of the best observed frontal face views are stored. Then similar faces belonging to different trajectories are clustered using a spectral clustering technique on the feature space constructed by 2DPCA.

Czirjek *et al.* proposed a method for automatically detecting human faces in generic video sequences [5]. For face detection, skin color filtering is carried out on a selected number of frames per video shot and projected into an eigenspace, the reconstruction error being the measure of confidence for presence/absence of face. Then the confidence score for the entire video shot is calculated. An incremental procedure using a PCA-based dissimilarity measure are employed in conjunction with spatiotemporal correlation to cluster extracted faces into a set of face classes.

Satoh proposed several face sequence matching methods and compared the performance of these methods by the accuracy of face sequence annotation [6]. The accuracy was evaluated using considerable amount of actual drama videos. The feature spaces were constructed by using Eigenface-based method, Fisher's linear discriminant-based method, subspace-based method, and kernel subspace-based method. Class information was utilized in those methods except Eigenface-based method.

But for automatic casting, we have to classify the face sequences without class informations. We have to construct feature space without class informations. In this paper, we compare several dimension reduction methods, such as PCA, LDA, and LPP in terms of automatic face sequence classification in movie videos.

3 Face Sequence Detection

The overview of the face sequence classification method is shown in Fig. 1. At first all faces are extracted from each frame of the given movie videos by applying the popular face detector proposed by Viola and Jones [1]. Then they are merged as a face sequences if the faces in the consecutive frames belong to the same shot and have similar size and location. Histogram of Oriented Gradients (HOG)[2] features are extracted from each face image in the sequences and they are used to compare the similarity of the face sequences. Feature space is constructed by using the several dimension reduction methods, such as PCA, LDA, LPP[3]. For classification of the face sequences, we simply define similarity measures for a pair of the feature vectors in the constructed feature space.

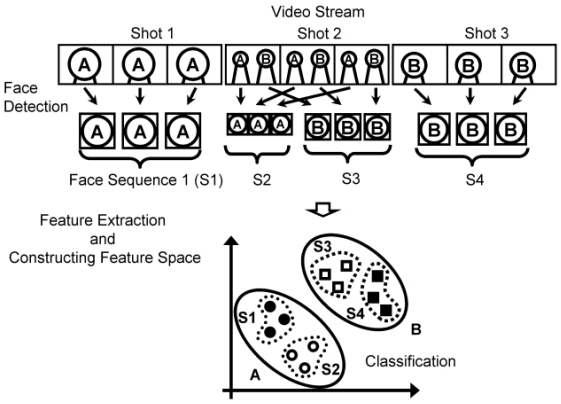

Fig. 1: The overview of face sequence classification method.

To extract face sequences from the given movie, we apply the popular face detector proposed by Viola and Jones (we call it *FD.V.J*)[1] to each frame in the movie. Then the extracted faces are merged as a face sequences if the faces in the consecutive frames belong to the same cut and have similar size and location.

3.1 Face detection

In this paper, the face detector proposed by *FD.V.J* is applied to detect faces in each frame of the given movie videos. It use a variant of Adaboost to select local rectangular features. A rectangular feature indicates difference of brightness between local rectangular regions neighboring each other. It can be efficiently computed at any scale and any location in the image by using a image representation called integral image. Each of the rectangular features is used as a simple base-classifier of Adaboost learning. In order to ensure fast classification, we have to exclude a large part of the available features and select a small subset of efficient features because the total number of rectangular features is very large. At each stage of Adaboost, a base classifier based on a rectangular feature is automatically selected from the all possible candidates. The selected classifiers

are combined to construct more complex classifier in a cascade manner. The cascade structure can increase the speed of the face detector by focusing attention on promising regions of the image. Also it is often possible to rapidly determine where a face might appear in an image. Due to an easy algorithm and high classification performance, their method of feature selection became popular and is used by many researchers for object detection. The face detector *FD.V.J* can operate in real-time and can detect nearly frontal faces at the different size and position.

3.2 Extraction of face sequences

Usually the size or the location of the captured person dose not change drastically in consecutive frames. Face images of the same person can be extracted by searching faces with similar size and location in the consecutive frames.

At first, the video is cut into continuous video shots. Then face images are detected from each frames of the video shot. The detected faces are compared in the adjacent frames. The two faces are regarded as the same if the face pair satisfy the following conditions:

- The size difference of the face pair is less than a specified threshold.
- The distance between the locations of the face pair is less than half of the face size.

By tracking the face pairs in the video shot, we can extract a sequence of faces. If more than one person appear in a given shot, several sequences are extracted. Since the current face detector is not perfect, sometimes it fail to detect the faces. So we rejected the face sequences in which the number of detected faces is extremely fewer than the number of frames in the video shot.

After a sequence of faces are extracted, face images are normalized to 20×20 pixels. Let $S_k = \{I_1^k, \cdots I_{n_k}^k\}$ be k-th face sequence, where I_i^k is the i-th normalized face image in the k-th sequence and n_k is the number of face images in the sequence. We denote a set of face sequences by $FS = \{S_1, \cdots S_K\}$ where K is the total number of face sequences in the video.

4 Feature Extraction

In the works by Dalal *et al.* [2], the Histogram of Oriented Gradients (HOG) features are extracted from all points on a dense grid in images. In this paper we use the grids of HOG features as the primitive features because they significantly outer perform existing feature sets for human detection as shown in [2]. Local object appearance and shape can often be characterized rather well by the distribution of local intensity gradients or edge direction. HOG features are calculated by taking orientation histograms of edge intensity in local region. HOG features are used in the SIFT descriptor proposed by Lowe [7].

In this paper, HOG features are extracted from 4 local regions with 16×16 pixels. At first, edge gradients and orientations are calculated at each pixel in this local region. Sobel filters are used to obtain the edge gradients and orientations. The gradient magnitude $m(x, y)$ and orientation $\theta(x, y)$ are calculated using the x- and y-directional gradients $dx(x, y)$ and $dy(x, y)$ computed by Sobel filter as

$$
\begin{aligned}
m(x,y) &= \sqrt{dx(x,y)^2 + dy(x,y)^2} \\
\theta(x,y) &= \begin{cases}
\tan^{-1}\left(\frac{dy(x,y)}{dx(x,y)}\right) - \pi \\
\text{if } dx(x,y) < 0 \text{ and } dy(x,y) < 0 \\
\tan^{-1}\left(\frac{dy(x,y)}{dx(x,y)}\right) + \pi \\
\text{if } dx(x,y) < 0 \text{ and } dy(x,y) > 0 \\
\tan^{-1}\left(\frac{dy(x,y)}{dx(x,y)}\right) \\
\text{otherwise}
\end{cases}
\end{aligned} \tag{1}
$$

This local region is divided into small spatial area called "cell". The size of the cell is 8×8 pixels. Histograms of edge gradients with 8 orientations are calculated from each of the local cells. Then the number of HOG features in the local region becomes $32 = 8 \times (2 \times 2)$ and they constitute a HOG feature vector. To avoid sudden changes in the descriptor with small changes in the position of the window, and to give less emphasis to gradients that are far from the center of the descriptor, a Gaussian weighting function with σ equal to one half the width of the descriptor window is used to assign a weight to the magnitude of each pixel. Since there are 4 local regions in a normalized face image, the total number of HOG features becomes $128 = 4 \times 32$. By extracting HOG features from each of the face sequence, we have a sequence of a HOG feature vectors $X_s = \{\boldsymbol{x}_1, \cdots \boldsymbol{x}_{n_s}\}$ for the given sequence s, where $\boldsymbol{x}_n$ is the feature vector extracted from the n-th normarilzed image in the face sequence.

5 Linear Feature Spaces

To evaluate the similarity of the face sequences, it is important to construct a proper features space in which the features in the face sequences of the same person becomes close and the features of the different persons becomes apart. In this paper, we compare the performance of the several dimension-

ality reduction methods, such as Principal Component Analysis (PCA), Linear Discriminat Analysis (LDA), and Locality Preserving Projection (LPP).

5.1 Principal Component Analysis

PCA is one of the well known techniques for dimensionality reduction. It has been applied to several computer vision problems. PCA can be defined as the orthogonal projection of the data onto a lower dimensional linear subspace, known as the principal subspace, such that the variance of the projected samples is maximized.

Let $X_{total} = \{\boldsymbol{x}_i | i = 1, \ldots, N\}$ be a set of all feature vectors extracted from a given video. We can apply PCA to this set of feature vectors. Then the principal scores are defined by using the projection matrix U_T as

$$\boldsymbol{y} = U_T^T(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \tag{2}$$

where the mean vector of the set of feature vectors are defined as $\bar{\boldsymbol{x}}_T = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i$. The optimum projection matrix U_T is obtained by solving the eigen equations of the total covariance matrix Σ_T

$$\Sigma_T U_T = U_T \Lambda, \quad (U_T U_T^T = I) \tag{3}$$

where the total covariance matrix Σ_T is defined as

$$\Sigma_T = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_T)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_T)^T. \tag{4}$$

This feature space reflects the variations of face appearances occurred in all feature vectors. We call this feature space PCA-ALL.

Another possibility to apply PCA to construct a feature space is to use the within class covariance matrix Σ_W instead of the total covariance matrix Σ_T in PCA-ALL. The within class covariance matrix is defined as

$$\Sigma_W = \frac{1}{N}\sum_{k=1}^{K} n_k \Sigma_k \tag{5}$$

where Σ_k is the covariance matrix of the k-th face sequence, n_k is the number of face images in the sequence. The optimum projection matrix U_W is obtained by solving the eigen equations of the total covariance matrix Σ_T

$$\Sigma_W U_W = U_W \Lambda, \quad (U_W U_W^T = I) \tag{6}$$

The constructed feature space reflects the variations within the face sequences. We call this feature space PCA-WITHIN.

The other possibility is to use the between class covariance matrix Σ_B instead of the total covariance matrix in PCA-ALL. The between class covariance matrix is defined as

$$\Sigma_B = \frac{1}{N}\sum_{k=1}^{K} n_k(\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}}_T)(\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}}_T)^T \tag{7}$$

where $\bar{x}_k$ is the mean vector of the k-th face sequence. The optimum projection matrix U_B is obtained by solving the eigen equations of the total covariance matrix Σ_T

$$\Sigma_B U_B = U_B \Lambda, \quad (U_B U_B^T = I) \tag{8}$$

The constructed feature space reflects the variations of the average vectors of each face sequence. We call this feature space PCA-BETWEEN.

5.2 Linear Discriminat Analysis

By assuming different sequences are different classes, we can also apply LDA to construct a feature space. The discriminant criterion $\boldsymbol{tr}(\hat{\Sigma}_W^{-1}\hat{\Sigma}_B)$ is used to evaluate the performance of the discrimination of the new features $\boldsymbol{y}$ and is maximized, where $\hat{\Sigma}_W$ and $\hat{\Sigma}_T$ are the within- and between-class covariance matrices defined on $\boldsymbol{y}$. The optimal coefficient matrix U_{LDA} is then given by solving the following eigen-equation

$$\Sigma_B U_{LDA} = \Sigma_W U_{LDA} \Lambda \quad (U_{LDA}^T \Sigma_W U_{LDA} = I). \tag{9}$$

In this feature space, it is expected that the features in the same face sequences becomes close and the features of the different sequences becomes apart.

5.3 Locality Preserving Projections

He *et al.* proposed a new linear dimensionality reduction algorithm called LPP [3]. It builds a graph incorporating neighborhood information of the data set. A transformation matrix which maps the data points to a subspace is constructed using the Laplacian of the graph. This transformation preserves local neighborhood information.

At first, the weight W_{ij} between the sample i and j in the data are calculated as

$$W_{ij} = \exp(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{t}) \tag{10}$$

where $t \in \boldsymbol{R}$ is a parameter. We denote the $N \times N$ symmetric weight matrix W whose elements is W_{ij}.

Then the optimal projection matrix are given by solving the following eigenvector problem

$$X_{total} L X_{total}^T U_{LPP} = X_{total} D X_{total}^T U_{LPP} \Lambda \tag{11}$$

where D is the diagonal weight matrix whose entries are column sums of W, $D_{ij} = \Sigma_j W_{ij}$, and

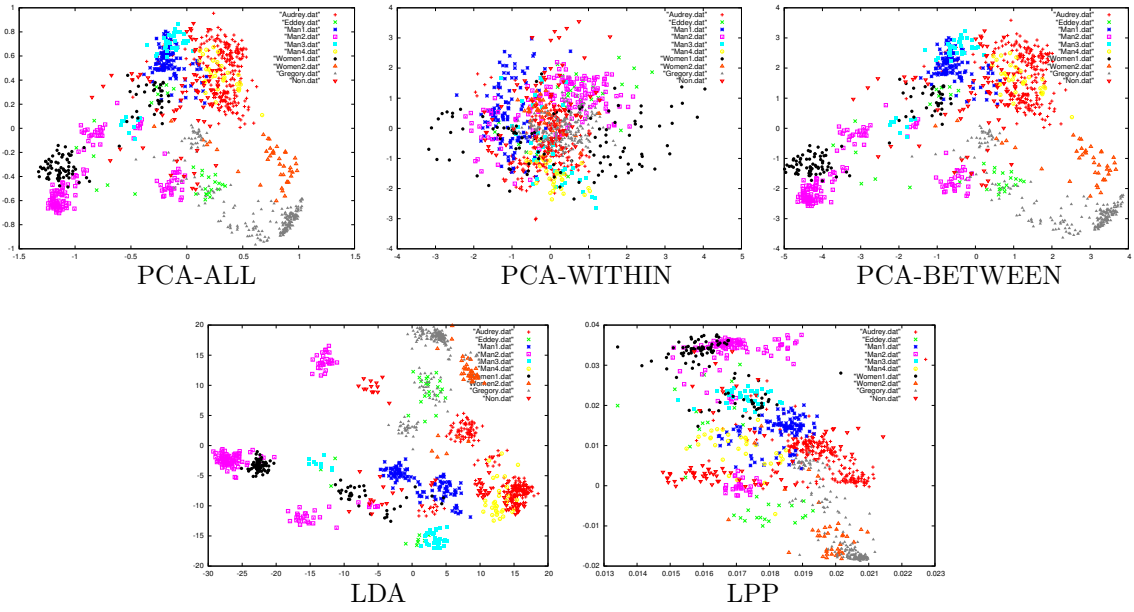

Fig. 2: Feature spaces constructed by dimensionality reduction methods PCA-ALL, PCA-WITHIN, PCA-BETWEEN, LDA, and LPP. The same person are denoted with the same color.

$L = D - W$ is the Laplacian matrix. Let $U_{LPP} = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_d]$ be the first d unitary orthogonal solution vectors Eq.(11), corresponding to the d smallest eigenvalues in the order of $0 \leq \lambda_1 \leq \cdots \leq \lambda_d$.

6 Face Sequence classification

To classify persons in videos, we must have a measure to evaluate the similarity between a pair of face sequences. Let $\{\boldsymbol{y}_1^p, \ldots, \boldsymbol{y}_{n_p}^p\}$ be a set of feature vectors extracted from a face sequence S_p.

One of the simplest measure of two set of feature vectors $\{\boldsymbol{y}_1^p, \ldots, \boldsymbol{y}_{n_p}^p\}$ and $\{\boldsymbol{y}_1^q, \ldots, \boldsymbol{y}_{n_q}^q\}$ is the Euclidian distance between the averages of each set. The measure is defined as

$$D(S_p, S_q) = ||\bar{\boldsymbol{y}}_p - \bar{\boldsymbol{y}}_q||^2, \tag{12}$$

where the averages are defined as $\bar{\boldsymbol{y}}_p = \frac{1}{n_p}\sum_{i=1}^{n_p} \boldsymbol{y}_i^p$ and $\bar{\boldsymbol{y}}_q = \frac{1}{n_q}\sum_{i=1}^{n_q} \boldsymbol{y}_i^q$.

Another similarity can be defined using the discriminant criterion as

$$J(S_p, S_q) = tr(\Sigma_T^{-1}\Sigma_B), \tag{13}$$

where Σ_T and Σ_B are the total covariance matrix and the between class covariance matrix of these two set of feature vectors. This measures the degree of separation of the two set of feature vectors.

Thus we can decide that the two sequences are obtained from the same person if the similarity is sufficiently high. Then two sequences are merged and construct a cluster.

7 Experimental Results

In the experiments, we used a set of face sequences extracted from the part of the popular movie entitled "Roman Holiday". The size of the video is 640×480. It's length is about 8 minutes with 7000 frames. The face images that are larger than 70×70 pixels are extracted from this video by face detector *FD.V.J.* By tracking faces in the video shots, 52 face sequences were extracted. Total 11 persons are included in the extracted face sequences. All the detected face images are resized to 20×20.

Then we compared the constructed feature space by the dimensionality reduction methods described in section 5. All face images in the extracted sequences were used to construct the feature spaces. Fig. 2 shows the constructed feature spaces. In this figure, all the faces in the extracted sequences are shown and the same person is plotted with the same color. It is notice that persons are well separated in the face spaces obtained by PCA-ALL, PCA-BETWEEN, LDA and LPP. Especially the distributions obtained by PCA-ALL and PCA-BETWEEN look like similar. On the other hand, the distributions of each person are confused in PCA-WITHIN. This means that the face space obtained by PCA-WITHIN is not suitable for classification of face sequences. In the feature space obtained by LDA, persons are well separated but there are clusters corresponding to each sequence in the same person. This means that the feature space ob-

Number of sequence

PERSON	A	B	C	D	E	F	G	H	I	J	K
Sequence	9	12	4	5	5	2	2	4	3	1	5

PCA-BETWEEN + Euclidian Distance

PERSON	A	B	C	D	E	F	G	H	I	J	K
Cluster1	0	2	1	4	3	1	0	4	0	0	1
Cluster2	1	8	1	0	0	0	0	0	3	0	0
Cluster3	8	0	0	0	0	0	0	0	0	0	0

LPP + Euclidian Distance

PERSON	A	B	C	D	E	F	G	H	I	J	K
Cluster1	0	8	0	0	0	0	0	0	1	0	0
Cluster2	4	0	0	0	0	0	0	0	0	0	0
Cluster3	3	0	0	0	0	0	0	0	0	0	0

PCA-BETWEEN + Discriminant Criterion

PERSON	A	B	C	D	E	F	G	H	I	J	K
Cluster1	7	4	1	0	0	0	0	0	2	0	1
Cluster2	0	0	0	5	0	0	0	2	0	0	2
Cluster3	0	5	0	0	0	0	0	0	0	0	0

LPP + Discriminat Criterion

PERSON	A	B	C	D	E	F	G	H	I	J	K
Cluster1	0	2	1	4	2	0	0	3	0	0	1
Cluster2	7	1	0	0	0	0	0	0	1	0	1
Cluster3	0	8	0	0	0	0	0	0	0	0	0

Table 1: The number of occurrences of each person in terms of the number of sequences in the 3 largest clusters obtained by the simple classification method.

tained by LDA is too sensitive to the sequences and is not suitable to evaluate the similarity between sequences.

To evaluate the goodness of the feature spaces constructed by these dimensionality reduction methods, we applied the simple classification method described in the section 6. All the face sequences were classified into several clusters. Then we identified the persons included in the 3 largest clusters. Table 1 shows the number of occurrences of each person in terms of the number of sequences in the 3 largest clusters. The table includes the results obtained in the feature spaces obtained by PCA-BETWEEN or LPP using Euclidian distance or discriminant criterion as measures for the similarity of face sequences. Feature dimensions were reduced from 128 dimensions to 13 dimensions by both methods. It is notices that the combinations of LPP and Euclidian distance give better results.

References

[1] P.Viola and M.Jones, "Rapid object detection using a boosted cascade of simple features." IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2001.

[2] N.Dalal and B.Triggs, "Histograms of Oriented Gradients for Human Detection" IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005.

[3] X.He and P.Niyogi, "Locality Preserving Projection" In Advances in Neural Information Processing Systems, 16, MIT press, 2004.

[4] S.Foucher and L.Gagnon, "Automatic detection and clustering of actor faces based on spectral clustering techniques," Porc. of Fourth Canadian Conference on Computer and Robot Vision (CRV'07), 2007.

[5] C.Czirjek, N.O'Connor, S.Marlow and N.Murphy, "Face detection and clustering for video indexing applications," Proc. of Advanced Concepts for Intelligent Vision Systems, September 2-5, 2003

[6] S.Satoh, "Comparative evaluation of face sequence matching for content-based video access," Proc. of the 4-th Int'l Conf. on Automatic Face and Gesture Recognition (FG2000), pp.163-168, 2000.

[7] D.G.Lowe "Distinctive Image Features from Scale-Invariant Keypoints" International Journal of Computer Vision (IJCV),60(2):91-110,2004.