ENHANCED MODEL SELECTION FOR MOTION SEGMENTATION

L. Zappella, X. Lladó and J. Salvi

Institute of Informatics and Applications, University of Girona, Girona, Spain

ABSTRACT

In this paper a novel rank estimation technique for trajectories motion segmentation within the Local Subspace Affinity (LSA) framework is presented. This technique, called Enhanced Model Selection (EMS), is based on the relationship between the estimated rank of the trajectory matrix and the affinity matrix built by LSA. The results on synthetic and real data show that without any a priori knowledge, EMS automatically provides an accurate and robust rank estimation, improving the accuracy of the final motion segmentation.

Index Terms— Machine Vision, Image Motion Analysis, Motion Segmentation

1. INTRODUCTION

The problem of dividing an image into background and foreground, also known as segmentation, is a crucial step for many computer vision applications. When the subject of the analysis is a video, rather than a still image, there is an additional information that can be exploited: the motion. In such a case the segmentation is also known as *motion segmentation*.

A great number of researchers have focused on the motion segmentation problem, however, despite the vast literature, performances of most of the algorithms still fall far behind human perception. A recent review on motion segmentation techniques can be found in [1]. Among feature based approaches, the Local Subspace Affinity (LSA) [2, 3] seems the most promising framework being able to deal with different types of motion: independent, rigid, articulated and non-rigid. Tron and Vidal concluded that LSA is the best performing algorithm (among LSA, GPCA and a RANSAC based approach) in case of non missing data [4].

One of the main problems of LSA is that it heavily relies on the rank estimation of the trajectory matrix. In the original proposal [2] this estimation was done by a model selection technique which requires the knowledge about the input sequence noise level in order to tune a sensitive parameter. Aiming to overcome this limitation we propose the Enhanced Model Selection (EMS) technique, a novel rank estimation for trajectory matrix, which does not require any tuning process nor a priori knowledge. EMS is based on the relationship between the estimated rank of the trajectory matrix and the affinity matrix built by LSA. EMS not only solves the previously explained limitation, but by improving the rank estimation also provides a more accurate motion segmentation.

In the next section the overview on the new rank estimation is given. The experiments obtained with synthetic and real sequences are presented in section 3. Finally, in section 4 conclusions are drawn.

This work has been supported by the Spanish Ministry of Science projects DPI2007-66796-C03-02 and DPI2008-06548-C03-03/DPI. L. Zappella is supported by the Catalan government scholarship 2009FI_B1 00068.

2. A NEW RANK ESTIMATION

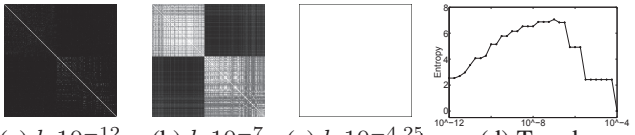
2.1. Local Subspace Affinity (LSA)

LSA is a framework for trajectories motion segmentation under affine projection proposed by Yan and Pollefeys [2, 3]. LSA can be summarized as follows:

1. build a trajectory matrix $W_{2f\times p}$, where f is the number of frames of the input sequence and p is the number of tracked feature points;
2. estimate the rank of $W_{2f\times p}$; this step is accomplished by a Model Selection technique (MS) inspired by the work of Kanatani [5]:

$$r = argmin_r \frac{\lambda_{r+1}^2}{\sum_{i=1}^{r} \lambda_i^2} + kr \qquad (1)$$

   being λ_n the n^{th} singular value of $W_{2f\times p}$, and k a parameter that depends on the noise of the tracked point positions: the higher the noise level is, the larger k should be [2];
3. project every trajectory, which can be seen as a vector in $\mathbb{R}^{2f}$, onto an $\mathbb{R}^r$ unit sphere;
4. estimate by singular value decomposition the subspace generated by each trajectory (and its nearest neighbours) in the new space;
5. compute an affinity matrix A, where the affinity between each pair of trajectories is the inverse of the distance between the generated subspaces;
6. cluster A in order to have the final motion segmentation.



(a) k:10^{-12} (b) k:10^{-7} (c) k:$10^{-4.25}$ (d) Trend

Fig. 1. (a-c): Affinity matrices of an input sequence computed after the MS rank estimation with different k values (black is minimum affinity, white is maximum affinity); (d): Entropy trend of the affinity matrices varying k value.

One of the weakest points of this framework, is the fact that the rank estimated by equation 1 requires the parameter k to be tuned depending on the input sequence noise level. Tuning k is a very important step. In fact, using a fixed k, or an improperly tuned k, may result in high misclassification rates [4]. The parameter k is so sensitive that Tron and Vidal, in their implementation of LSA, avoid the MS and fix the new space size to $4n$, where n is the number of motions. In this way two new assumptions are made: rigid motion (the theoretical maximum rank for a rigid motion is 4 [6]), and knowledge of the number of motions n. The aim of EMS is to provide automatically an accurate rank estimation of the trajectory matrix looking for the best k value, without requiring any knowledge or making any assumptions.

2.2. Estimated rank and affinity matrix relationship

EMS finds automatically a good k value exploiting the relationship between the rank of $W_{2f\times p}$ estimated by MS (LSA step 2) and the computed affinity matrix A (LSA step 5). Such relationship can be seen in Fig. 1(a) to 1(c), where the affinity matrices of an input sequence with two motions (maximum rank 8) are shown. When the rank of $W_{2f\times p}$ is estimated using an inappropriate k value, the affinity matrix does not provide any useful information. Specifically, if k is too small MS tends to overestimate the rank and from the affinity matrix it is possible to infer only that every trajectory is independent from every other. This is the case of Fig. 1(a), where a $k = 10^{-12}$ leads to a rank of 57. The opposite happens when k is too high and MS tends to underestimate the rank of $W_{2f\times p}$, in this case from the affinity matrix it is possible to infer only that every trajectory is strongly related to any other. This is the case of Fig. 1(c), where $k = 10^{-4.75}$ leads to a rank of 3. On the other hand, when k is well tuned the rank estimation tends to be closer to the real rank of $W_{2f\times p}$, and the affinity matrix can be used for a successful segmentation. This is the case of Fig. 1(b), where $k = 10^{-7}$ leads to a rank of 8.

Therefore, if a measure of the quality of the affinity matrix is found, it would be possible to evaluate the accuracy of the estimated rank. In such a case the rank estimation and the affinity matrix computation could be repeated iteratively until a "good" affinity matrix, hence an accurate rank estimation, is obtained.

2.3. Entropy of the affinity matrix

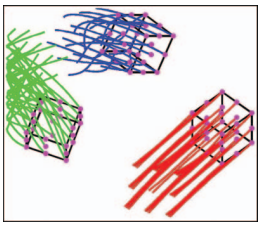
In order to find a measure able to describe the quality of the affinity matrix it is necessary to define what a "good" affinity matrix is. Ideally, in presence of at least two motions, a perfect affinity matrix would have only two values: the highest possible value, for every pair of trajectories that belong to the same motion, and the lowest possible value, for every pair of trajectories that belong to different motions. However, due to noise and dependent motions the affinity matrix rarely has only two values. Most frequently, it has two modes close, but not necessarily equal, to the highest and to the lowest possible value. The two peaks of the modes correspond to those pairs of sequences clearly related, or clearly unrelated. In addition there is a certain amount of in between values for those pairs that are somehow related but not completely similar. In contrast, bad affinity matrices are those that do not differentiate enough between related and unrelated trajectories which means that the histogram of those matrices is unimodal with a mode corresponding to very high or very low values.

The trends of different statistical parameters, extracted from the affinity matrices obtained going from overestimation to underestimation of the rank, have been analysed. From this study it emerged that the entropy defined as:

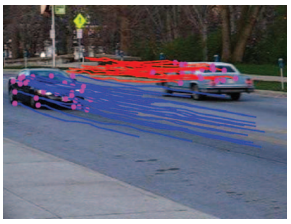
$$E(A) = -\sum(I\ log_2(I)) \tag{2}$$

where I contains the histogram counts of A, can be used as a measure of the quality of the affinity matrix. In fact, when the rank of $W_{2f\times p}$ is overestimated or underestimated, the corresponding affinity matrices are homogeneous respectively with low values and high values, whereas if the rank estimation is accurate the affinity matrix contains a wider range of values. Fig. 1(d) shows the trend of the entropy computed on the affinity matrices of the same sequence showed in Fig. 1(a) to 1(c). Entropy starts with low values when $k = 10^{-12}$, as in Fig. 1(a). As k increases, and A tends to the one of Fig. 1(b), the entropy also increases and reaches its maximum when $k = 10^{-7}$. After that point the entropy starts decreasing and it drops to zero when A becomes completely homogeneous, as in Fig. 1(c).

Hence, a "good" affinity matrix could be built using the rank estimation of $W_{2f\times p}$ that led to the maximum entropy. However, building all the affinity matrices going from small to high k values is computationally expensive. Nonetheless, the entropy trend has a property that can be exploited in order to speed up the maximum entropy localization: entropy trend has only one global maximum and no local maxima nor minima. This happens because going from overestimation to underestimation of the rank of $W_{2f\times p}$, the space size onto which the trajectories are projected decreases. Every time the space size becomes smaller the distance between every trajectory subspace tends to decrease. However, the distance between trajectory subspaces that belong to the same motion decreases faster than the distance between trajectory subspaces that be-



(a) Synthetic sequence



(b) Real sequence

Fig. 2. Two frames of input sequences used to test EMS.

long to different motions. If the dimension keeps decreasing, eventually any subspace will be close to any other, to the point that all the trajectories will lie on exactly the same subspace. Summarizing, going from small to high k values the distance between subspaces tends to decrease, therefore the affinity tends to increase until it reaches the maximum value. This is the reason why the entropy trend of the affinity matrix is a convex function without local minima or maxima. Hence, it is possible to exploit the gradient of the entropy trend in order to have a good approximation of where the maximum entropy is, drastically reducing the amount of calculations.

3. RESULTS

In order to evaluate EMS we compare the results obtained using LSA with MS[1] and our implementation of LSA with EMS (available at `http://eia.udg.edu/~zappella`). Both algorithms provide the final segmentation applying Spectral Clustering [7] to the affinity matrix as suggested in [2]. We perform experiments on synthetic sequences, as in Fig. 2(a), and real sequences, as in Fig. 2(b).

3.1. Synthetic experiments

The synthetic database is composed of video sequences of 30 frames with rotating and translating cubes, where each one has 26 tracked points evenly spaced on its surface. There are 5 sequences with random motion with 2, 3, 4, and 5 motions, for a total of 20 sequences. This set is then perturbed with random gaussian noise with a standard deviation of 0.5, 1, 1.5, 2, 2.5 and 3 pixels, composing a synthetic database with a total of 140 sequences. The misclassification rate of LSA with MS presented in this section is obtained after a tuning step choosing among different k values the one that led to the lowest average misclassification rate ($k = 10^{-7.5}$). EMS did not require any tuning process.

Fig. 3(a) shows the boxplot of the misclassification rate averaged over all the synthetic sequences. The average misclassification is much lower for the EMS version of LSA: 1.1% against 5.6%. It should be noticed how broader the MS first and the second quartile ranges are compared with EMS ones. Moreover, the number of outliers (in terms of misclassification rate of the sequences) is considerably smaller with EMS. Furthermore, the highest EMS misclassification is only 10% while for MS the highest misclassification is 60%.

The misclassification rates depending on the noise level and on the number of motions are shown in Fig. 3(b) to 3(d), while Fig. 3(e) shows the misclassification averaged among all number of motions. The trend with only 2 motions is not shown as both algorithms have a low misclassification rate (less than 1%) independently from the noise level. With 3 and 4 motions MS misclassification remains low as long as the noise level has a standard deviation lower than 1.5 pixels. After this noise level the misclassification increases dramatically. With 5 motions the misclassification rate starts becoming considerably high even with a noise standard deviation of only 0.5 pixels. From these results the sensitivity of MS about the relationship between the k value and the noise level is confirmed, but another problem arises: k seems to be influenced not only by the noise level but also by the number of motions. On the other hand, LSA with EMS misclassification rate remain more stable either when the noise level increases and when the number of motions increases. The average misclassification never rises above 4% (see Fig. 3(e)).

3.2. Real experiments

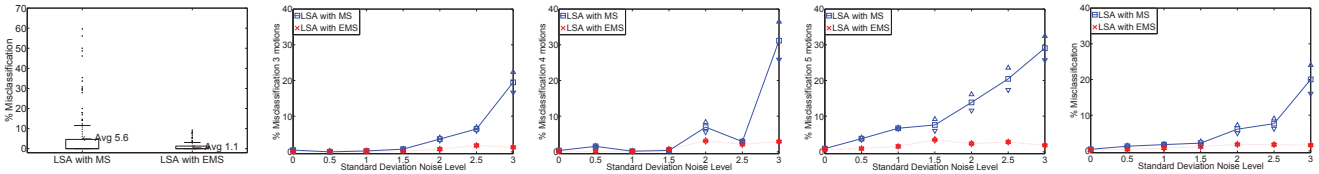
In order to test EMS also with real sequences we use the Hopkins155 database[1] [4], which is a reference database for motion segmentation, composed of 155 real video sequences: 120 with 2 motions and 35 with 3 motions. Again, for LSA with MS we computed the misclassification rate using different k values and we are presenting in this section the lowest average misclassification (obtained with $k = 10^{-7}$).

Fig. 4(a) shows the boxplot of the misclassification rate. As in the synthetic results, EMS always has a lower misclassification rate and more compact quartile ranges. These results prove that EMS always provides a better rank estimation of $W_{2f\times p}$, and it does so in an automatic fashion.

Inside the Hopkins155 database there are different types of sequences: checkboards, traffic and articulated/non-rigid. The checkboard is the main group, 104 videos, hence it is likely that the type and the amount of noise does not change much as most of the sequences are taken in the same environment. For the purpose of testing the EMS with bigger noise changes, we created another six databases derived from the Hopkins155 adding random gaussian noise, with standard deviation of 0.5, 1, 1.5, 2, 2.5 and 3 pixels, to the tracked point positions. The original database plus the six derived from it compose a bigger database with 1085 video sequences.

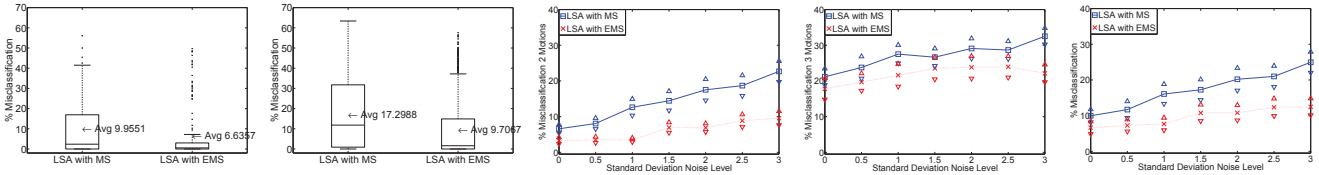
We compared again LSA with MS using $k = 10^{-7}$ and LSA with EMS. Fig. 4(b) shows the boxplot of the misclassification rate on the modified Hopkins155 database. As before, EMS has lower average misclassification and more compact quartile ranges. As expected, the increment in the misclassification rate (from Fig. 4(a) to 4(b)) is bigger with MS than

[1] Available at `http://www.vision.jhu.edu`



(a) Synthetic (b) 3 Motions (c) 4 Motions (d) 5 Motions (e) Average

Fig. 3. Synthetic Sequences. (a): Misclassification boxplots; (b-d): mean (□, ×) and variance (△, ▽) trends of the misclassification rate with different number of motions; (e): mean (□, ×) and variance (△, ▽) trends of the misclassification rate averaged overall the number of motions.



(a) Hopkins155 (b) Hopkins155 with noise (c) 2 Motions (d) 3 Motions (e) Average

Fig. 4. Real Sequences. (a-b): Misclassification boxplots; (c-d): mean (□, ×) and variance (△, ▽) trends of the misclassification rate with different number of motions; (e): mean (□, ×) and variance (△, ▽) trends of the misclassification rate averaged overall the number of motions.

with EMS, MS increment is more than double than the EMS one: 7.3% against 3.1%.

Fig. 4(c) to 4(e) show the average misclassification and variance for each algorithm changing the noise level and the number of motions. Both algorithms have more problem to deal with 3 motions, but also in this case EMS has a better behaviour. From these plots it is also possible to evaluate the divergence between MS and EMS misclassification. Misclassification increases for both algorithms when the noise level rises but EMS is able to contain the misclassification better than MS. Considering two and three motions averaged together (see Fig. 4(e)) the difference of the misclassification between MS and EMS goes from 3.32% without any added noise to 12.22% with 3 pixels of standard deviation noise.

4. CONCLUSION

In this paper a novel EMS rank estimation for trajectory matrices has been presented. EMS exploits the relationship between the trajectory matrix rank and the affinity matrix built by LSA. The results on synthetic and real sequences proved that EMS provides a more accurate rank estimation leading to a more successful motion segmentation. Moreover, standard MS requires some tuning process regarding the noise level of the input sequence and the number of motions in order to provide low misclassification rate, while EMS is able to adapt automatically without any a priori knowledge. As far as we know, this is the first automatic LSA. In fact, until now the rank of $W_{2f\times p}$ was either estimated assuming the knowledge of the amount of noise [2, 3], or assuming the knowledge about the number and the type of motions [4].

5. REFERENCES

[1] L. Zappella, X. Lladó, and J. Salvi, "Motion segmentation: A review," in *Frontiers in Artificial Intelligence and Applications*, 2008, vol. 184, pp. 398–407.

[2] J. Yan and M. Pollefeys, "A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate," in *Lect. Notes Comput. Sc.*, 2006, vol. 3954, pp. 94–106.

[3] J. Yan and M. Pollefeys, "A factorization-based approach for articulated nonrigid shape, motion and kinematic chain recovery from video," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 30, no. 5, pp. 865–877, 2008.

[4] R. Tron and R. Vidal, "A benchmark for the comparison of 3-d motion segmentation algorithms," *Proc. CVPR IEEE*, pp. 1–8, 2007.

[5] K. Kanatani, "Motion segmentation by subspace separation and model selection," *IEEE I. Conf. Comp. Vis.*, vol. 2, pp. 586–591, 2001.

[6] Carlo Tomasi and Takeo Kanade, "Shape and motion from image streams under orthography: a factorization method," *Int. J. Comput. Vision.*, vol. 9, no. 2, pp. 137–154, 1992.

[7] Jianbo Shi and Jitendra Malik, "Normalized cuts and image segmentation," in *IEEE Trans. Pattern Anal. Machine Intell.*, 2000, pp. 888–905.