Overview of algorithms for face detection and tracking

Nenad Markuš

***Abstract*—The human face is central to our identity. It plays an essential role in everyday interaction, communication and other routine activities. Thus, face detection and tracking algorithms are of great importance for human-machine interaction. It is an extensively studied field and a large body of research has been done. This paper gives a a broad overview of basic principles and some representative methods for face detection, tracking and facial feature extraction.**

***Index Terms*—face tracking, facial feature localization, face detection**

I. Motivation

The human face is central to our identity. It plays an essential role in everyday interaction, communication and other routine activities. Detection and tracking the face and its features thus potentially opens a very wide range of applications. Using the face as a means of human-computer interaction is helping disabled people improve their daily lives, and may become a hands-free alternative in other applications or an entertaining element in innovative games. Model-based coding of facial video relies on facial tracking to enable very low bitrate video communication [3]. As specified in the MPEG-4 International Standard [38] it enables full facial communication at less than 10 kbit/s. Similarly, tracking facial actions is a basis for driving computer animated faces in games and entertainment applications. Machine lip reading has been used to enhance speech recognition [40]. In the field of virtual reality, view-dependent rendering relies on head tracking to dynamically generate the correct perspective on a 3D scene depending on the location of the user, creating a sensation of 3D on regular 2D screens. Knowing the position of the head is also an essential component of some types of autostereoscopic displays. Many gaze estimation attempts can be greatly simplified if the system knows an approximate locations of facial features [24]. In security and authentication applications, face tracking is a front-end for facial recognition and may also be used for liveness detection to avoid fraud by substituting live face with an image. In augmented reality applications, face tracking is a basis for augmenting the face with additional graphics, enabling commercial applications such as virtual try-on of eyewear, hairstyles or makeup, as well as games and artistic creations. Facial tracking is increasingly finding use in safety applications to detect situations such as sleepiness or lack of attention while driving or using hazardous machinery. There is also a growing interest toward emotional intelligence in human-computer interaction paradigms. In order to react appropriately to a human, the computer needs to have the perception of the emotional state of the human. It has been asserted in [43] that the most informative channel for machine perception of emotions is through facial expressions in video streams. Last but not least, facial tracking is a tool for analysing facial motion in various research fields, such as chewing analysis within food-releated research. All these applications create a strong motivation for research and development of fast and robust methods for face detection and tracking.

Face detection and tracking is an extensively studied field and a large body of research has been done on these subjects. This paper gives a short overview of basci principles and some representative methods for face detection, tracking and facial feature extraction. We do not give a comprehensive survey of the history of face tracking.

II. Face detection

Face detection in still images is an important problem in machine vision and is often the first step in many vision-based applications that involve interaction between a human and a machine. It can be formulated as follows. Given an arbitrary image, the goal of face detection is to determine whether or not there are any faces in the image and, if present, return the location and extent of each face.

In their seminal work [47], Viola and Jones introduced their object detection framework that can be applied to human faces. At runtime, the algorithm is capable of processing images extremely rapidly — real-time frame rates can be achieved on consumer hardware. It uses a boosted [19] cascade of simple classifiers based on rectangular Haar-like features that can be computed rapidly. The algorithm has a learning stage and both high detection rate and low false positive rate can be achieved with a sufficiently large training set. Although the algorithm can be trained to detect various objects, it was motivated primarily by the problem of face detection and high quality cascades are publicly available.

A detailed survey of recent advances in face detection can be found in [50].

III. Active appearance models in face modeling

Active appearance model (AAM) is an algorithm for matching a statistical model of a deformable object to a new image. It was introduced by Cootes [12]. As this method can be used to generate a variety of object instances photorealistically, it soon proved to be effective in a variety of applications. Some include general object tracking [41], red-eye removal [48], gait analysis [31], medical image segmentation [34], as well as face tracking and facial expression recognition.

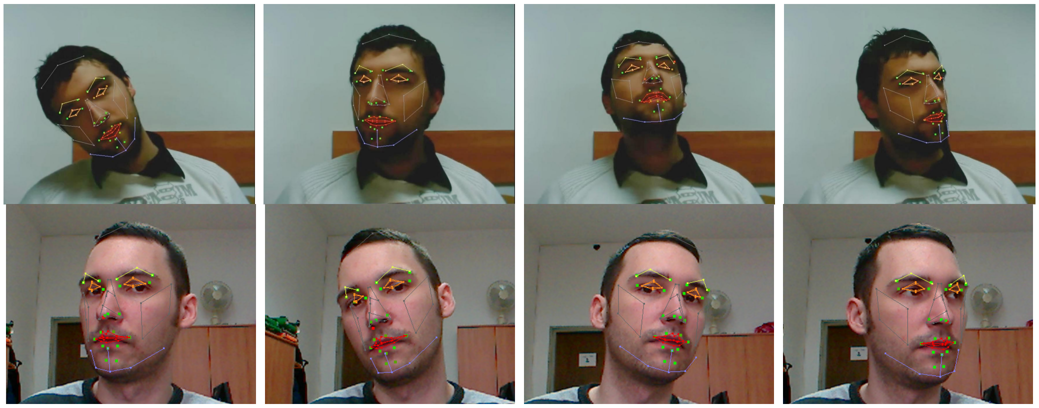

Fig. 1. A face tracker in action.

The rest of this section gives a short introduction to the basic principles of AAMs and describes some applications in face tracking, facial feature detection and expression anaylsis.

A. Basic AAM principles

1) Modeling: AAMs are built during a training phase. The training phase requires a set of images with labeled landmarks. These images have to be provided in advance by an expert. The key idea of AAMs is to statistically model the texture and the shape of a deformable object.

The shape, represented by a vector formed by concatenating the positions of labeled landmarks, is first normalized using Procrustes analysis [22] and then projected on a low dimensional subspace using PCA. Subsequently, based on the corresponding points, images in the training set are warped to the mean shape to produce shape-free patches. The texture from these patches is sampled and placed in a vector. Again, the dimensionality is reduced using PCA.

This compressed representation greatly simplifies the fitting of an AAM to a new image as there are a lot less parameters to optimize.

2) Model fitting: Once we have created an AAM, it is of interest to optimize the parameters and fit the model to an object appearing in a new image. These parameters enable us to obtain information such as object rotation and translation, landmark positions, etc. However, the task of AAM fitting is an unconstrained optimization problem, which is often difficult to solve. It is usually adresssed using a gradient descent algorithm or linear regression.

When the AAM method is applied to tracking, recognition and image synthesis, successful results can be obtained even when perspective, rotational and translational transformations are present. However, problems arise in the presence of nonlinear variations of texture and shape, fast motion and other special circumstances. Many improvements to the basic AAM algorithm have been proposed over the years to compensate for these deficiencies. These new approaches have been recently described in [21]. The paper gives a comprehensive study of AAM algorithms with respect to efficiency, discrimination and robustness. It can serve as a guide for further research in the field.

B. Applications to face modeling

AAMs have been extensively used in face modeling and are a standard tool for near-frontal face tracking, facial feature detection, expression analysis and face synthesis. Some examples are given in the text that follows.

1) Tracking: A very efficient face tracker has been described in [45]. The main contribution of this work is the demonstration of an active appearance model that runs in real-time on a mobile device. The high performance can be attributed to fast model fitting using the nonlinear regression framework decribed in [51] and an efficient implementation (for example, the use of fixed-point arithmetics and nearest neighbor texture sampling). It is unclear how well can this tracker cope with large appearance and pose variations. This system proves that AAMs can be very efficient if implemented properly.

2) Classification: Edwards *et al.* [17] used an AAM for face recognition. In their approach, the face is located in an image using AAM search and the corresponding parameters are extracted. The shape and texture parameters are then used as input to a classification algorithm. This approach is theoretically robust to confounding factors such as lighting, pose and expression variation. The hypothesis is partially confirmed in their paper.

Abboud *et al.* [1] applied and AAM for expression recognition. Their system uses LDA in the training phase to extract the most discriminative features according to the seven basic human expressions: anger, disgust, fear, joy, surprise, sadness and neutral. Euclidean and Mahalanobis distance are employed for classification of the expression in novel images.

3) Synthesis: Similar to the linear projection from the appearance to the parameter space, AAM is able to reconstruct faces from the parameters [29]. This approach has been used by various researchers to generate face portraits [11] and facial expressions [1], [44], [32].

IV. 3D MODEL-BASED FACE TRACKING

Face tracking can be formulated as a problem of fitting a rigid or non-rigid 3D model to the face of the user.

A. Tracking using a rigid model

This subsection describes some trackers that use a rigid 3D face model. These form the basis for more advanced face trackers.

A general motion-based approach to face tracking is based on the estimation of displacements of pixels from one frame to another. The displacements might be estimated using optical flow methods, block-based motion estimation methods or motion estimation in a few image patches only, giving a few motion vectors only but at very low computational cost. The estimated motion field is used to compute the motion of the object model using some optimization method (extended Kalman filtering, least squares, Nelder-Mead downhill simplex optimization, etc.). The object model is used only for transforming the 2D motion vectors to 3D object model motion. The problem with such methods is the drifting or the long sequence motion problem: it has been observed that trackers of this kind accumulate motion errors and eventually lose track of the face. Examples of such trackers can be found in [5], [7]. Some other motion-based trackers [15], [16], [33] add various model-based constraints to improve performance and combat drift.

Instead of using a generic tracking framework, some systems include the appearance of the face with the 3D model. Systems of this kind sample the face texture in the first frame and formulate face tracking as an image registration problem (old frames are considered only to constrain the search space). Several approaches have been developed to robustly track faces under large pose variations. They rely on a rigid 3D face/head model, which can be a cylinder [10], [49], an ellipsoid [36], or a mesh [46], [42]. The model is fit to the image by matching either local features or facial texture. These kind of trackers do not suffer from drifting. Instead, problems arise when the model is not flexible enough to cope with the situation in the new frame. This can especially be true when rapid changes in facial expressions are present.

B. Tracking using a non-rigid model

Non-rigid face tracking algorithms build on the work described in previous subsection. The idea is to add an additional set of parameters to the 3D model in order to allow deformations of the mesh (these often include lip or eyelid movement, configuration of the distance between the eyes or the length of the nose, etc.).

Candide [2] is a simple wireframe face model (figure 2). Ever since its creation, it has been a popular face model in many research labs. This can be attributed to its simplicity and public availability. The model contains a 3D description of a face as well as parameters for controlling facial expressions and face shape variation between individuals. The geometry is determined by the 3D coordinates of the vertices in a model-centered coordinate system. To modify the geometry, the model implements a set of parameters. These parameters can be divided into two groups. The first group consists of action parameters. These define the dynamic changes in the facial geometry, like eye blinking and lip movement. The second group of parameters are called shape parameters. The shape parameters control the static deformations that cause individuals to differ from each other. A common startegy is to fix the shape parameters at the beginning of the tracking session and to dynamically estimate the action parameters. The rest of this section describes some face trackers that us the Candide face model.

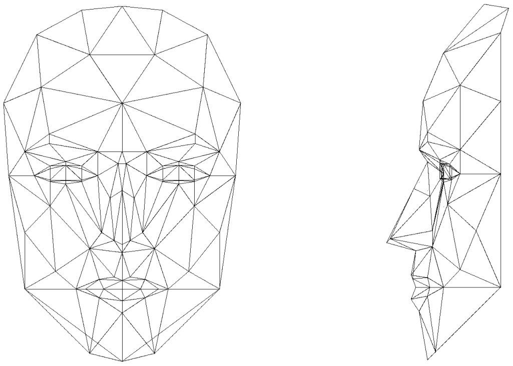

Fig. 2. The Candide face model.

The main idea of Ström [42] was to construct a rigid textured 3D model of the face from the first frame of the video stream and to update its motion based on the displacements of a few patches. The patches are extracted from the 3D textured model and tracked in the video frames using normalized cross-correlation. The global motion of the face is reconstructed using an extended Kalman filter and the structure from motion framework [4]. The tracker was later extended by Ingemars and Ahlberg [26] to include the non-rigid motion of the model, and thus track facial action. In other words, the action parameters of the Candide model are estimated along with the face rotation and translation parameters. The tracking process recursively alternates between two steps. In the first step, the displacements of blocks of pixels are determined and used to estimate the global motion of the face, i.e. it is purely motion-based. The second step is exactly the same as in Ström's approach and is used to combat the drift problem.

In [30], the authors have presented a system based on tracking characteristic points on the face and intensity features sampled from the facial texture. Additionally, the authors extended the Candide model in order to deal with challenging face poses: they added additional polygons to cover the head sides. This was motivated by the fact that the side of the face has rich appearance information. The displayed tracking quality was very good, as was demonstrated on the Boston University Face Tracking [10] dataset. The authors have reported that the system was capable of processing three frames per second on a PC. This makes it inapropriate for implementation on mobile devices.

V. Decision forests in face modeling

Randomized forests [9] have recently become popular among machine vision researchers and engineers due to their simplicity, efficiency and performance capabilities. A randomized forest is a structure for regression or classification that consists of an ensemble of decision trees [8]. The output is given by averaging the outputs of all the trees in the ensemble. At each node of each tree, a binary test is computed[1] and the incoming image is sent recursively to the left or right child node until a leaf is reached. Leaf nodes contain a simple model that encodes the output of the tree. One of the main reasons behind the success of randomized forests is due to the randomness present in the learning process: randomness is injected in the induction of each single tree by bootstrap sampling the training set and randomly choosing features as split candidates at each node (this procedure reduces the correlation between different trees in the ensemble and leads to the reduction of generalization error [23]). Applications include fast keypoint recognition [37], joint localization from depth cameras [39], regression in medical imaging [13], age estimation [35], object tracking, detection and recognition [20], as well as head rotation and facial feature extraction.

Belle *et. al.* [6] presented a system for detecting and classifying faces in images in real-time based on randomized classification forests. First, the forest is trained to distinguish faces and non-faces within training images. In the second step, the forest is trained to distinguish between individual people. This yields an efficient face recognition system — it is capable of face recognition in real-time due to the use of Haar-like features in binary tests at each tree node. The obtained face detector performed comparatively to the Viola-Jones detector available in OpenCV albeit with a higher false positive rate. It was noted that the SVM outperformed the randomized forest in the face classification task but at the cost of much higher training time and a much higher time required for the classification.

Fanelli *et. al.* [18] addressed the problem of head pose estimation from depth images. They decided to use randomized regression forest. The developed system is capable of estimating the 3D coordinates of the nose tip and the rotation of the face. An important thing they noticed is that learning the regression forest on rather generic face patches requires enormous amount of training data in order to achieve accurate estimates. Thus, they synthesized a big training set using a statistical model of the human face and demonstrated that this approach yields good results in practice (this was also observed in [39]).

Dantone *et. al.* [14] build on the framework of [13], [18] and develop a real-time facial feature detector. The system is based on regression forests that learn the relationships between facial image patches and facial feature locations (the initial rectangle containing a face is obtained using a Viola-Jones detector). It is capable of impressive performance on the challenging Labeled Faces in the Wild [25] database (an example can be seen in figure 3). The authors have not demonstrated the system's performance on mobile devices. Thus, it is not clear weather the algorithm is suitable for face tracking in these environments.

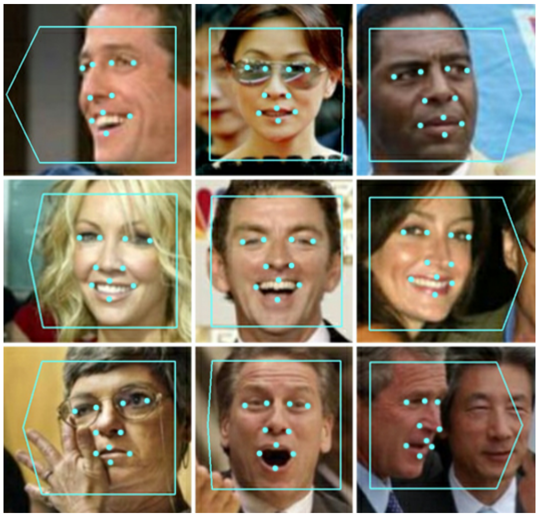

Fig. 3. The performance of the system described in [14] on the Labeled Faces in the Wild database [25].

Kalal *et. al.* [27] developed a system for long-term tracking of a human face in unconstrained videos based on Tracking-Learning-Detection [28] framework. The framework is based on the fact that an efficient classifier (randomized forest) can be used to represent the decision boundary between the object (for example, a face) and its background. The forest can be learned on-line, during real-time performance, using structural constraints in common video streams, such as the fact that the face trajectory is continous (of course, the whole system needs to be initialized manually or with a Viola-Jones detector). The resulting system is resistant to occlusions and appearance changes. It successfully tracked and detected a face in a video sequence with over 35000 frames.

VI. Conclusion

The paper gives a broad overview of face tracking, detection and facial feature localization algorithms. It can serve as a starting point for researchers that want to get quickly into the field.

References

[1] B. Abboud, F. Davoine, and M. Dang. Facial expression recognition and synthesis based on an appearance model. *Signal Process., Image Commun.*, 2004.

[2] J. Ahlberg. Candide-3 – an updated parametrised face. Technical report, Linköping University, 2001.

[3] J. Ahlberg and R. Forchheimer. Face tracking for model-based coding and face animation. *International journal of imaging systems and technology*, 2003.

[4] A. Azerbayejani and A. Pentland. Recursive estimation of motion, structure, and focal length. *IEEE PAMI*, 1995.

[1] For example, this test could be a comparison of some Haar-like feature response with a predefined treshold.

[5] S. Basu, I. Essa, and A. Pentland. Motion regularization for model-based head tracking. In *CVPR 1996*, 1996.

[6] V. Belle, T. Deselaers, and S. Schiffer. Randomized trees for real-time one-step face detection and recognition. 2008.

[7] M. J. Black and Y. Yacoob. Recognizing facial expressions in image sequences using local parameterized models of image motion. *IJCV*, 1997.

[8] L. Breiman. *Classification and regression trees*. Chapman and Hall, 1984.

[9] L. Breiman. Random forests. *Journal of Machine Learinng Research*, 2001.

[10] M. L. Cascia, S. Scarloff, and V. Athitsos. Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3d models. *IEEE PAMI*, 2000.

[11] H. Chen, Z. Liu, C. Rose, Y. Xu, H. Y. Shum, and D. Salesin. Example-based composite sketching of human portraits. In *Int. Symp. Nonphotorealistic Anim. Rendering*, 2004.

[12] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE PAMI*, 2001.

[13] A. Criminisi, E. Konukoglu, and J. Shotton. Regression forests for efficient anatomy detection and localization in ct studies. In *Medical Computer Vision 2010: Recognition Techniques and Applications in Medical Imaging, MICCAI workshop*, 2010.

[14] M. Dantone, J. Gall, G. Fanelli, and L. Van Gool. Real-time facial feature detection using conditional regression forests. 2012.

[15] D. DeCarlo and D. Metaxas. Optical flow constraints on deformable models with applications to face tracking. *IJCV*, 2000.

[16] F. Dornaika and J. Ahlberg. Face and facial feature tracking using deformable models. *International Journal of Image and Graphics*, 2004.

[17] G. J. Edwards, T. F. Cootes, C.J. Taylor, and Manchester M Pt. Face recognition using active appearance models. Technical report, University of Manchester, 1998.

[18] G. Fanelli, J. Gall, and L. V. Gool. Real time head pose estimation with random regression forests. 2011.

[19] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting, 1997.

[20] J. Gall, A. Yao, N. Razavi L. V. Gool, and V. Lempitsky. Hough forests for object detection, tracking, and action recognition. *IEEE PAMI*, 2011.

[21] X. Gao, Y. Su, X. Li, and D. Tao. A review of active appearance models. *IEEE Systems, Man, and Cybernetics*, 2010.

[22] C. R. Goodall. Procrustes methods in the statistical analysis of shape. *J. Roy. Statist. Soc. B*, 1991.

[23] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Science and Business, 2009.

[24] T. Heyman, V. Spruyt, and Alessandro Ledda. 3d face tracking and gaze estimation using a monocular camera. In *Proceedings of the 2nd International Conference on Positioning and Context-Awareness*, 2011.

[25] G. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts, Amherst, 2007.

[26] N. Ingemars and J. Ahlberg. Feature-based face tracking using extended kalman filtering. In *Swedish Symposium on Image Analysis*, 2007.

[27] Z. Kalal, K. Mikolajczyk, and J. Matas. Face-tld: Tracking-learning-detection applied to faces. *ICIP*, 2010.

[28] Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking-learning-detection. *IEEE PAMI*, 2012.

[29] A. Lantis, C. J. Taylor, and T. F. Cootes. Automatic interpretation and coding of face images using flexible models. *IEEE PAMI*, 1997.

[30] S. Lefvre and J.-M. Odobez. View-based appearance model online learning for 3d deformable face tracking. In *Int. Conf. Computer Vision Theory and Applications*, 2010.

[31] X. Li, S. Maybank, S. Yan, D. Tao, and D. Xu. Gait components and their application to gender recognition. *IEEE SMC C*, 2008.

[32] I. Macedo, E. V. Brazil, and L. Velho. Expression transfer between photographs through multilinear aams. In *Comput. Graph. Image Process.*, 2006.

[33] M. Malciu and F. Preteux. A robust model-based approach for 3d head tracking in video sequences. In *International Conference on Face and Gesture Recognition*, 2000.

[34] S. C. Mitchell, B. P. Lelieveldt, R. J. van der Geest, H. G. Bosch, J. H. Reiber, and M. Sonka. Multistage hybrid active appearance model matching: segmentation of left and right ventricles in cardiac mr images. *IEEE Trans Med Imaging*, 2001.

[35] A. Montillo. Age regression from faces using random forests. In *16th IEEE International Conference on Image Processing*, 2009.

[36] L.-P. Morency, J. Whitehill, and J. Movellan. Generalized adaptive view-based appearance model: Integrated framework for monocular head pose estimation. In *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2008.

[37] M. Özuysal, P. Fua, and V. Lepetit. Fast keypoint recognition in ten lines of code. In *IEEE Conference on Computing Vision and Pattern Recognition*, 2007.

[38] I. S. Pandžić and R. Forchheimer. *MPEG-4 Facial Animation: The standard, implementations and applications*. John Wiley and Sons, 2002.

[39] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. Technical report, Microsoft Research, 2011.

[40] P. L. Silsbee and A. Bovik. Computer lipreading for improved accuracy in automatic speech recognition. *IEEE Speech and Audio Processing*, 1996.

[41] M. B. Stegmann. Object tracking using active appearance models. In *Proc. Danish Conf. Pattern Recog. Image Anal.*, 2001.

[42] J. Ström. Model-based real-time head tracking. *EURASIP Journal on Applied Signal Processing*, 2002.

[43] Y. Sun, N. Sebe, M. S. Lew, and T. Gevers. Authentic emotion detection in real-time video. *Computer Vision in Human-Computer Interaction*, 2004.

[44] B. Theobald, I. A. Matthews, J. F. Cohn, and S. M. Boker. Real-time expression cloning using active appearance models. In *Int. Conf. Multimodal Interfaces*, 2007.

[45] P. A. Tresadern, M. C. Ionita, and T. F. Cootes. Real-time facial feature tracking on a mobile device. *IJCV*, 2011.

[46] L. Vacchetti, V. Lepetit, and P. Fua. Stable real-time 3d tracking using online and offline information. *IEEE PAMI*, 2004.

[47] P. Viola and M. Jones. Robust real-time object detection. *IJCV*, 2001.

[48] J. Wan, X. Renm, and G. Hu. Automatic red-eyes detection based on aam. In *IEEE Int. Conf. SMC*, 2004.

[49] J. Xiao, T. Moriyama, T. Kanade, and J. F. Cohn. Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3d models. *IEEE PAMI*, 2000.

[50] C. Zhang and Z. Zhang. A survey of recent advances in face detection. Technical report, Microsoft Research, 2010.

[51] S. K. Zhou, B. Georgescu, X. S. Zhou, and D. Comaniciu. Image based regression using boosting method. In *ICCV*, 2005.