Go With The Flow: Hand Trajectories in 3D via Clustered Scene Flow

Simon Hadfield and Richard Bowden

Centre for Vision Speech and Signal Processing,
University of Surrey, Guildford, Surrey, UK, GU2 7XH
{S.Hadfield,R.Bowden}@surrey.ac.uk

Abstract. Tracking hands and estimating their trajectories is useful in a number of tasks, including sign language recognition and human computer interaction. Hands are extremely difficult objects to track, their deformability, frequent self occlusions and motion blur cause appearance variations too great for most standard object trackers to deal with robustly.

In this paper, the 3D motion field of a scene (known as the Scene Flow, in contrast to Optical Flow, which is it's projection onto the image plane) is estimated using a recently proposed algorithm, inspired by particle filtering. Unlike previous techniques, this scene flow algorithm does not introduce blurring across discontinuities, making it far more suitable for object segmentation and tracking. Additionally the algorithm operates several orders of magnitude faster than previous scene flow estimation systems, enabling the use of Scene Flow in real-time, and near real-time applications.

A novel approach to trajectory estimation is then introduced, based on clustering the estimated scene flow field in both space and velocity dimensions. This allows estimation of object motions in the true 3D scene, rather than the traditional approach of estimating 2D image plane motions. By working in the scene space rather than the image plane, the constant velocity assumption, commonly used in the prediction stage of trackers, is far more valid, and the resulting motion estimate is richer, providing information on out of plane motions. To evaluate the performance of the system, 3D trajectories are estimated on a multi-view sign-language dataset, and compared to a traditional high accuracy 2D system, with excellent results.

1 Introduction

The 3D trajectories of freely moving hands within sign language, provide valuable information for the recognition of signs. However tracking hands is extremely difficult using standard object trackers, as they move and deform rapidly, and frequently occlude each other. In this paper, a three dimensional motion field is estimated along with associated structure. Trajectories are then extracted by clustering, both in the estimated structure and motion field.

As the hands can change appearance so rapidly, even adaptive object detectors are generally unable to reliably localise them. Instead, previous approaches operate by segmenting the input image, according to some weak knowledge of hand regions such as skin colour and motion assumptions. These segmentation results are then processed to extract a small number of smooth regions, which are then labelled. The sequence of positions associated with each label is then the estimated trajectory.

In this paper, the segmentation system is replaced with a scene flow estimator. Rather than grouping segmented pixels into regions based on proximity, the estimated

motion field is then clustered in 6D space $(x, y, z, \Delta x, \Delta y, \Delta z)$ before labelling. The advantage of this approach is that there is less chance of the hand regions merging in the 6D space, than on the 2D image plane (hands may have distinct motion directions, or be separated in depth). Additionally, the use of Scene Flow naturally provides hand trajectories in the 3 world dimensions, rather than the 2D projection onto the image plane, generally estimated in other approaches.

The scene flow estimation algorithm used in this paper, employs a collection of weighted hypotheses similar to a particle filter. These hypotheses represent structure and motion estimates in the scene, which are filtered based on appearance information from multiple viewpoints, to estimate the overall scene characteristics.

In brief, the novelties of this paper are:

- Hand trajectory estimation based on Scene Flow.
- Estimation of 3D hand trajectories, rather than in the image plane.
- Extending the Scene Particle algorithm [7] to incorporate additional task specific information.

1.1 Related Work

Hand Tracking The most accurate approaches to hand tracking involve the user wearing special gloves. In [12] data gloves incorporating accelerometers are used to directly obtain hand position. To avoid the expense of such systems, authors such as Kadir et al. [11] use vision techniques, coupled with gloves of specific colours, and ensure these colours are not present elsewhere in the scene. However, gloves are encumbering for users, particularly in sign language applications. Many authors [10, 1, 8] thus employ skin segmentation techniques to estimate hand position. Hand trajectories estimated from skin segmentation, suffer from confusion between the two hands and the face. Han et al. [8] attempt to deal with this using a Kalman filter, and a simple motion model for the 3 objects.

The current state-of-the-art in hand tracking is the system proposed by Buehler et al. [4]. This combines colour segmentation with Histogram of Oriented Gradient based limb appearance and pictorial structure, to perform robust 2D tracking. Even so, the approach still requires repeated processing, back and forth through the video to resolve occlusions and is therefore extremely slow.

Scene Flow Estimation Scene flow is the dense, 3 dimensional motion field of a scene, of which optical flow is a 2D projection. Estimation of Scene flow is generally formulated as an optimisation problem, with a global energy function over a reference image, estimating the X,Y and Z motion at every pixel [9]. This approach tends to be extremely computationally complex, mitigated only by specialist GPU implementations [17].

Basha et al. [2] showed that estimating the motion field in 3D scene space, rather than the image plane, improves performance. This is due to improved validity of assumptions such as local smoothness, and allows easy extension to any number of views.

A regularisation cost is generally used to constrain the optimisation. This tends to cause over-smoothing at discontinuities in the motion field, such as object boundaries. Wedel et al. [20] among others, attempt to reduce this effect by basing the regularisation strength on image gradients. Another approach employed by Li et al. [13] involves segmenting the image, and applying regularisation within segments.

Several approaches [14, 6, 5] represent the problem as sparse trajectory estimation, similar to the application tackled in this paper. However the trajectories correspond to the motion of scene mesh vertices, requiring initialisation, and limiting the range of possible motions.

The approach proposed in [7] is inspired by particle filtering rather than optimisation, making it less computationally expensive, easily parallelisable, and supporting multiple-hypotheses. Additionally, the regularisation of the energy function is avoided, allowing more accurate estimation at discontinuities, which is invaluable for object segmentation and tracking tasks such as 3D hand tracking. In addition we propose a number of modifications to the original formulation which improve performance in typical HCI scenarios.

1.2 Paper Structure

The following section (2.1) provides an overview on the use of scene flow estimation for extracting 3D trajectories. An overview of the probabilistic formalisation for multiview scene flow estimation is presented in section 2.2. The integration of the skin likelihood model is explained in section 3. Then, the details of the Scene Particle algorithm (4) and a forwards & backwards extension (4.1) are presented. Section 4.2 describes the clustering of the motion field. In section 5.1 the proposed scene flow algorithm is compared with alternative techniques for 3D motion field estimation. Results of trajectory estimation in sign language sequences are presented in section 5.2, with comparisons to existing 2D techniques. Finally discussion of the approach, with possibilities for future investigation, are presented in section 6.

2 Proposed Framework

2.1 Trajectory Estimation

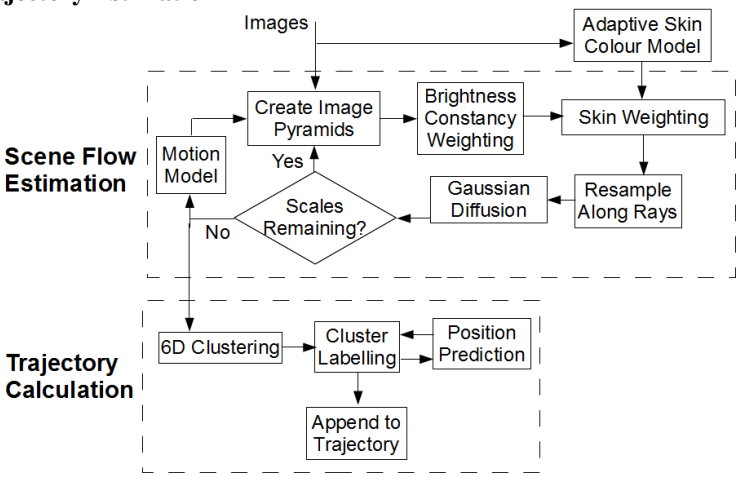

Fig. 1. Overview of trajectory estimation architecture.

Figure 2.1 provides an overview of the proposed approach. The scene flow estimation system processes the input data from the multi-camera rig, and produces an estimate of 3D scene structure, along with an estimated 3D motion field. Additional input

is provided to the scene flow estimator from a skin colour model, in order to concentrate estimation on relevant portions of the scene.

Clusters are then generated from the estimated structure and motion field. These clusters are labelled, based on the predicted positions and velocities from the previous frame. Finally the labelled positions are appended to the trajectories, and a constant velocity motion model is used to estimate the cluster positions.

2.2 Multi-View Probabilistic Scene Flow

Estimating the structure and motion of an unconstrained scene, is equivalent to finding structure points $\mathbf{r}$ which are most likely to be objects, and their associated 3D motion $\mathbf{v}$. This can be viewed as a 6 dimensional, continuous, probability distribution, from which we wish to extract a set of peaks.

Based on an observation $\mathbf{i}$, it is possible to represent the posterior probability $\mathbf{p}(\mathbf{r}, \mathbf{v}|\mathbf{i})$ as the product of the prior probability and the likelihood.

$$\mathbf{p}(\mathbf{r}, \mathbf{v}|\mathbf{i}) \propto \mathbf{p}(\mathbf{i}|\mathbf{r}, \mathbf{v})\mathbf{p}(\mathbf{r}, \mathbf{v}) \tag{1}$$

As the distribution is across scene positions and velocities, the prior probability $\mathbf{p}(\mathbf{r}, \mathbf{v})$ is easily obtained from the posterior probability at the previous time instant, in conjunction with a motion model.

The likelihood $\mathbf{p}(\mathbf{i}|\mathbf{r}, \mathbf{v})$ is calculated using the Brightness Constancy Assumption. This assumption is frequently used for motion and structure estimation, and states that the colour of a world point is identical, when viewed from any direction, and over time. When a scene is viewed by a set of M cameras in any multi-view setup, with projection matrices $\mathit{\Pi}_{1..M}$, images $I_{1..M}$ are obtained. Over a period of T frames, the image sequences $I_{1..M}^{1..T}$ are generated as the observation $\mathbf{i}$. The Brightness Constancy Assumption implies that any true world point $\mathbf{r}$, with colour vector $\mathbf{u}$, moving between frames with a velocity $\mathbf{v}$, will satisfy the condition:

$$(\forall q)\,, I_q^t\left(\mathit{\Pi}_q \mathbf{r}\right) = I_q^{t-1}\left(\mathit{\Pi}_q\left(\mathbf{r} - \mathbf{v}\right)\right) = \mathbf{u} \tag{2}$$

In human interaction scenarios, generally the lighting of the scene, and sometimes even the response characteristics, vary across cameras. To improve the validity of the brightness constancy assumption used for scene flow estimation, the images are normalised to the same mean and variance, before processing. Thus we can assume that the colour $\mathbf{u}$ of a point, is the average of the colours in each (normalised) image. This means the square of the deviation from the equation 2 becomes the variance of colours across all images (at both the previous and current frame).

Using this, the likelihood distribution is calculated as in Equation 3, where the function $var(\mathbf{r}, \mathbf{v})$ is the variance of the projection colours, and $\epsilon = 0.001$. This provides a smooth approximation of L_1 (see [3]). Projections are performed with subpixel accuracy, and the resulting colours are obtained using bilinear interpolation.

$$\mathbf{p}(\mathbf{i}|\mathbf{r}, \mathbf{v}) = \frac{\mathbf{1}}{\sqrt{\mathbf{var}(\mathbf{r}, \mathbf{v})^{\mathbf{2}} + \epsilon^{\mathbf{2}}}} \tag{3}$$

3 Skin Colour Model

For the estimation of hand trajectories, additional task specific information is included in the system, in the form of an extra term in the likelihood equation 3. This term relates to the likelihood that each combination of $\mathbf{r}$ and $\mathbf{v}$ belong to an object of interest, based on an adaptive skin colour model. The purpose of this is to concentrate scene flow estimation into regions of interest, allowing for faster computation.

The skin colour model comprise of 2 probability distributions in RGB colour space, one for skin colour ($\mathbf{P_s^q}$) and one for background colour ($\mathbf{P_{bg}^q}$). For a point in the scene space $\mathbf{r}, \mathbf{v}$, with colour vector $\mathbf{u}$, the extra likelihood term $\mathbf{h(i|r,v)}$ is calculated using the log-likelihood ratio of the skin and background models, as shown in equation 4. By averaging the responses across viewpoints, the system obtains resistance to single view occlusions, as objects of interest will still exhibit strong responses in the remaining views. A limit α on the responses is estimated from training data.

$$\mathbf{h(i|r,v)} = \sum_{\mathbf{q=1}}^{\mathbf{M}} \mathbf{min}\left(\mathbf{log_{10}}\left(\frac{\mathbf{P_s^q(u)}}{\mathbf{P_{bg}^q(u)}}\right), \alpha\right) \tag{4}$$

The models are all bootstrapped from a generic colour model, and then allowed to adapt to the data based on the results of face detection. The colour distribution within the central region of the face detection is combined with $\mathbf{P_s}$, and the distribution across the remainder of the image is added to $\mathbf{P_{bg}}$.

4 Scene Particle Algorithm

Distributions across all possible $\mathbf{r}$ and $\mathbf{v}$ are obviously intractably large, and so the posterior $\mathbf{p}(\mathbf{r}, \mathbf{v}|\mathbf{i})$ is represented by a collection of N weighted points. These points are referred to as Scene Particles. Each particle p_n includes a 3D position in the world $\mathbf{r}$, and a 3D motion vector $\mathbf{v}$, and also has an associated weight w_n.

As the space is continuous, multiple Scene Particles are maintained along the same epipolar ray, with different depths. In addition, multiple Scene Particles may exist at the same structure position, with different motion hypotheses.

A well known problem (see [19]) with particle filtering systems that operate over long periods of time, is that eventually the particles converge to the largest mode of the distribution. If only the single "best" point in the distribution is needed, as in standard particle filtering, this is acceptable. However, in the Scene Particle algorithm, we are not interested in the most probable point, but instead in the set of all high probability points. To prevent many Scene Particles collecting onto a small number of structure points, leaving other areas of the scene probability space under-explored, the weight w_n^t of scene particle p_n at time t, is taken as the average of it's previous weight w_n^{t-1} and it's likelihood given new observations $\mathbf{i}$, as in equation 5. The effect of this change, is that likelihoods based on more recent observations, have an increased effect on the current weighting, resulting in better coverage being maintained over long sequences.

$$w_n^t = \frac{\mathbf{p(i|p_n)} + \mathbf{w_n^{t-1}}}{\sum_{j=0}^{N}\left(\mathbf{p(i|p_j)} + \mathbf{w_j^{t-1}}\right)} \tag{5}$$

The Scene Particle population is resampled as in standard particle filtering. This process serves to concentrate hypotheses in regions of the 6D space with high weighting, more thoroughly exploring them, while reducing the number of Scene Particles used to examine low weighted regions. To ensure structure and motion is estimated densely across the scene a technique is used, referred to here as Ray Resampling. In this approach, the particles along every epipolar ray are resampled separately. This is equivalent to using a separate particle filter per ray, while allowing particles to move between filters at frame transitions.

For estimation of sign language trajectories, only a sparse motion field is required. For this reason, rays containing no Scene Particles are left empty. However, for the comparison with other scene flow techniques (section 5.1), empty rays are populated by random selection from neighbouring rays, enabling complete coverage, and reflecting neighbouring distributions.

An iterative estimation approach is used, to allow the estimated motion field time to converge. New images from every viewpoint are converted to multi-scale image pyramids and estimation is performed in a coarse to fine manner as in [9, 16], this helps particles avoid local minima during convergence. Additionally several inner iterations are performed within each level of the pyramid, successively reducing the diffusion of the motion model. This allows Scene Particles to transition from exploratory behaviour, to more precise estimation. Experiments in this paper are performed using 3 pyramid levels, with 3 noise loops, for a total of 9 iterations per frame.

4.1 Forwards & Backwards Brightness Constancy

Equation 3 specifies the likelihood of a Scene Particle, based on brightness constancy in every viewpoint, at the current and previous frame. An additional assumption common in tracking systems, is that of constant velocity, assuming the time between frames is small. By employing this assumption, we extend the likelihood calculation to include the future frame, giving the new condition shown in equation 6. As in section 2.2 the deviation from this condition, is the variance of a Scene Particles projected intensities across all viewpoints at all 3 frames (previous, current and future).

$$(\forall q), I_q^t(\Pi_q \mathbf{r}) = I_q^{t-1}(\Pi_q(\mathbf{r}-\mathbf{v})) = I_q^{t+1}(\Pi_q(\mathbf{r}+\mathbf{v})) = \mathbf{u} \tag{6}$$

Examining Scene Particles in additional images improves the chances of disambiguating between similar motions. However, a future frame is required for the variance calculation, necessitating a 1 frame delay in estimation. In the remainder of the paper, scene flow based on this extended condition is referred to as “Forwards/Backwards” (FB) scene flow.

4.2 Clustering & Labelling

After the structure and motion field have been estimated, clustering is performed via expectation maximisation, with three clusters. Clustering is performed in the 6 dimensional space of location and motion. Cluster centres are initialised by prediction from previous results, using a constant velocity motion model. At frame 0, where there is no previous prediction, clusters are initialised on the top, bottom left and bottom right extremities of the scene. After the output of the Scene Flow estimation is clustered, labels are assigned to the new cluster centres, in order to create consistent trajectories for the 3 objects (head and 2 hands).

5 Results

5.1 Scene Flow Comparisons

In order to perform a quantitative evaluation of the proposed Forwards/Backwards extension to the Scene Particle algorithm, tests were run on a standard scene flow estimation dataset. The dataset is available from vision.middlebury.edu [18], and was originally intended for multi-view stereo reconstruction. It consists of a set of 9 rectified images from 9 cameras viewing a static scene for a single frame. It is possible to simulate a scene in which all objects are in motion, viewed from multiple cameras, by taking a number of images to act as frame 0 in each view. The remaining images can then be used as the second frame of each sequence.

As ground truth disparity (with subpixel accuracy) is provided for the scene, the motion between frames is known. This allows performance to be measured from 4 viewpoints, utilising 8 of the images for standard 2 frame estimation. When testing the FB (3 frame) approach from section 4.1, 3 viewpoints are simulated using all 9 images. In order to operate on general data rather than focus on hand trajectories, the adaptive skin colour model is disabled during these tests.

Table 1 contains the comparisons. The Root Mean Square Error (RMSE) is measured for each estimated value, averaged across the scene. The Average Angular Error (AAE) of the Scene Flow is also measured as in [20]. Each algorithm operates with a different number of views and, in the case of the FB extension, a different number of frames per view, the total amount of image data utilised is also displayed.

Algorithm	Dataset	Images Used	Optical Flow	Z Flow	Structure	AAE (deg)
Single View+Depth[7]	Teddy	4	0.11	0.00	-	5.04
Multiview	Teddy	8	1.12	0.07	1.10	4.13
Multiview+FB	Teddy	9	0.73	0.06	1.02	4.21
Single View+Depth[7]	Venus	4	0.08	0.00	-	5.50
Multiview	Venus	8	0.97	0.02	0.96	4.19
Multiview+FB	Venus	9	0.51	0.03	0.93	4.32
Single View+Depth[7]	Cones	4	0.09	0.00	-	5.10
Multiview	Cones	8	1.61	0.08	1.61	4.17
Multiview+FB	Cones	9	0.59	0.07	1.65	4.22

Table 1. Performance of Scene Flow multiview estimation compared to estimation using a single appearance and depth sensor as in [7]. Also results are shown including the Forwards/Backwards extension.

For these tests, only 10 Scene Particles per ray were used, leading to a runtime of less than a minute per frame for a non-parallelised implementation, which is several orders of magnitude faster than alternative scene flow techniques [9]. In the multiview scenario, motion estimation errors increase significantly, due to the need to estimate both motion and structure simultaneously. However, the additional viewpoints allow similar motions, which were previously ambiguous to be distinguished, leading to an increase in directional accuracy. Additionally the multiview approach does not require depth sensors, and can be applied in a wider range of scenarios.

The Forwards/Backwards confirmation approach leads to a significant improvement in motion estimation accuracy roughly halving optical flow errors, and providing

smaller improvements to structure and out of plane motion estimates. However, due to the need for additional frames, the FB approach uses fewer viewpoints than the standard multiview approach. This leads to increased difficulty distringuishing between similar motions, and consequently a small reduction in directional accuracy (similar to the gain found going from single view to multi-view).

5.2 Sign Language Trajectory Estimation

With the inclusion of the Trajectory calculation elements, and the adaptive skin colour model, the approach was then applied to sign language data, for hand trajectory estimation.

The 3D position ground truth for hands during sign language, is difficult to obtain accurately without the use of data gloves. The authors are not aware of any available natural sign language dataset, with such ground truth. Thus, results were obtained on newly captured multi-view dataset, for which 2D trajectories in the image plane were available. Figure 2 shows example frames from parts of this dataset. The previous 5 positions are displayed for the head and both hand trajectories. As can be seen, estimated 3D trajectories of all 3 objects are consistent with all viewpoints.

Figure 2 parts A-C are taken from a narrow baseline, 2 view, sequence. It can be seen that estimated 3D trajectories are consistent with the 2D observations from both viewpoints. However, figure 2.C illustrates a failure case, where the narrow baseline creates some ambiguity in the Z dimension. The resulting trajectory matches the data when projected to the left image plane, however the projection in the right viewpoint shows an erroneous 1 frame jump towards the face. Wider baseline systems such as Figure 2.D do not suffer from these errors. Additionally the third viewpoint in this sequence allows the system to operate robustly in the case of occlusions. Despite being completely occluded in the side view for several frames, the estimated trajectory for the right hand is consistent with the observations from the other 2 views.

Object	Agreement	X RMS error	Y RMS error
Head	100%	0.057	0.097
Right Hand	93.535%	0.191	0.100
Left Hand	88.054%	0.277	0.091

Table 2. Agreement between projection of estimated 3D trajectories, and 2D trajectories from an alternative system (values in palm widths).

Table 2 gives the comparison of the 3D trajectory to the extremely accurate 2D tracking approach from [15]. Trajectories were assumed to be in agreement if their separation was less than one third of a palm width. The percentage of frames where trajectories are in agreement is listed for each object, as is the RMS distance between trajectories in terms of palm widths. The close agreement between the trajectories indicates that the estimates are plausible, if the system was innacurate, it would be unlikely they would coincide so frequently in the frontal view. The results from the table provides a lower baseline on performance, as in some cases 2D tracking may be lost due to frontal occlusion while the 3D tracking is maintained by the other views. Results were compared over a 30,000 frame sequence. The greater accuracy of the head estimation is to be expected as it moves little during the sequences. The accuracy of the left hand is slightly higher than the right, as one viewpoint is placed on this side, and so the hand is occluded less often.

Fig. 2. (A) to (C) are taken from the narrow baseline, 2 view, sequence, showing the projection of 3D trajectories which are consistent with both viewpoints. (C) Illustrates a failure case where the narrow baseline of the 2 views gives rise to ambiguity in Z. (D) Shows a frame from the wide baseline, 3 view, sequence, illustrating robustness to object occlusions.

6 Conclusions

The results presented in this paper have shown that clustered Scene Flow is an effective technique for trajectory estimation tasks. Such approaches enable the estimation of trajectories in 3D rather than the image plane (providing invaluable information for recognition tasks), but spatially close objects moving in different directions to be well separated, reducing the crossing of hand tracks often observed in sign language tasks. In addition, the multi-view particle based approach allows propagation of information between frames, and the maintenance of multiple hypotheses at every point, improving robustness to occlusions.

Also in this paper, extensions of the Scene Particles algorithm have been demonstrated, to incorporate common hand tracking assumptions, such as small acceleration and skin colour probability, and to account for variations in illumination and camera responses between viewpoints.

In the future, the Scene Particle algorithm may be improved, by introducing smoothness constraints on the motion and structure field. Insofar as possible, such constraints should preserve object discontinuities, to maintain the advantages of the approach. The trajectory estimation system may be improved by investigating alternative clustering methods, and introducing a re-initialisation system when the distance between motion model prediction and cluster centre is too great.

Acknowledgements This work is supported by the European Community's Seventh Framework Programme (FP7/2007-2013) grant agreement no 231135 (Dicta-Sign).

References

1. Awad, G., Han, J., Sutherland, A.: A unified system for segmentation and tracking of face and hands in sign language recognition. In: ICPR. Hong Kong, China (Aug 2006)
2. Basha, T., Moses, Y., Kiryati, N.: Multi-view scene flow estimation: A view centered variational approach. In: CVPR (2010)
3. Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High accuracy optical flow estimation based on a theory for warping. In: ECCV (2004)
4. Buehler, P., Zisserman, A., Everingham, M.: Learning sign language by watching tv (using weakly aligned subtitles). In: CVPR. Miami, FL, USA (Jun 20 – 26 2009)
5. Courchay, J., Pons, J., Monasse, P., Keriven, R.: Dense and accurate spatio-temporal multi-view stereovision. In: ACCV (2009)
6. Furukawa, Y., Ponce, J.: Dense 3d motion capture from synchronized video streams. In: CVPR (2008)
7. Hadfield, S., Bowden, R.: Kinecting the dots: Particle based scene flow from depth sensors. In: ICCV. Barcelona, Spain (6-13 Nov 2011)
8. Han, J., Awad, G., Sutherland, A.: Automatic skin segmentation and tracking in sign language recognition. IET Computer Vision (2009)
9. Huguet, F., Devernay, F.: A variational method for scene flow estimation from stereo sequences. In: ICCV (2007)
10. Imagawa, K., Lu, S., Igi, S.: Color-based hands tracking system for sign language recognition. In: FGR. Nara, Japan (Apr 14 – 16 1998)
11. Kadir, T., Bowden, R., Ong, E., Zisserman, A.: Minimal training, large lexicon, unconstrained sign language recognition. In: BMVC (2004)
12. Kim, J.H., Thang, N.D., Kim, T.S.: 3-d hand motion tracking and gesture recognition using a data glove. In: ISIE (2009)
13. Li, R., Sclaroff, S.: Multi-scale 3d scene flow from binocular stereo sequences. In: Workshop on Motion and Video Computing (2005)
14. Neumann, J., Aloimonos, Y.: Spatio-temporal stereo using multi-resolution subdivision surfaces. IJCV (2002)
15. Pitsikalis, V., Theodorakis, S., Vogler, C., Athena, R., Maragos, P.: Advances in phonetics-based sub-unit modeling for transcription alignment and sign language recognition. In: Workshop on Gesture Recognition (2011)
16. Pons, J., Keriven, R., Faugeras, O.: Multiview stereo reconstruction and scene flow estimation with a global image-based match score. IJCV (2007)
17. Rabe, C., Müller, T., Wedel, A., Franke, U.: Dense, robust, and accurate motion field estimation from stereo image sequences in real-time. In: ECCV (2010)
18. Scharstein, D., Szeliski, R.: High-accuracy stereo depth maps using structured light. In: CVPR (2003)
19. Sidenbladh, H.: Probabilistic Tracking and Reconstruction of 3D Human Motion in Monocular Video Sequence. Ph.D. thesis, Stockholm Royal Institute of Technology (2001)
20. Wedel, A., Brox, T., Vaudrey, T., Rabe, C., Franke, U., Cremers, D.: Stereoscopic scene flow computation for 3d motion understanding. IJCV 95(1), 29–51 (2011)