Monocular Omnidirectional Head Motion Capture in the Visible Light Spectrum

Jeroen Lichtenauer and Maja Pantic
Department of Computing, Imperial College London
London, UK
http://ibug.doc.ic.ac.uk/research/head-motion-capture/

Abstract

Conventional marker-based optical motion capture methods rely on scene attenuation (e.g. by infrared-pass filtering). This renders the images useless for development and testing of machine vision methods under natural conditions. Unfortunately, combining, calibrating and synchronising a system for motion capture with a separate camera is a costly and cumbersome task. To overcome this problem, we present a framework for efficient, omnidirectional head-pose initialisation and tracking in the presence of missing and false positive marker detections. As such, it finally enables easy, accurate and synchronous head-motion capture as ground truth with or input for other machine vision algorithms.

1. Introduction

Accurate and reliable motion capture is essential to many applications in the field of robotics and natural human behaviour analysis. Of the many different methods for motion capture, the marker-free, passive, computer-vision-based approaches appear to be the most practical, as they do not require a person to wear any special clothing, markers or other equipment [9]. However, there is still a lot of progress to be made in the accuracy and robustness of passive computer-vision-based motion capture. Accurate ground truth pose information is essential for the development of motion capture methods. Especially when it comes to proper evaluation of performance [5]. Furthermore, with a method to obtain head pose robustly and accurately, other methods that may benefit by this information - such as for facial expression analysis - can already be developed and tested more easily. This ground truth has to be obtained using a sufficiently accurate alternative motion capture system.

When capturing naturalistic data for development of unobtrusive computer-vision-based motion capture methods, the method to be used for obtaining the ground truth obviously has to work in such a naturalistic environment. Ideally, the same camera will be used for obtaining the ground truth motion as well as for capturing naturalistic recordings of the target behaviour. Not only does this save the cost of extra hardware, it also implicitly solves the difficult problem of accurate spatial calibration and time synchronisation in case two separate data capture systems are used. Unfortunately, optical marker-based methods are aversely affected by the visual clutter in naturalistic environments. In existing motion capture systems, this is either solved by letting the subjects wear black clothing and by using black backgrounds, or by using markers that emit or reflect infrared light together with cameras that filter out the visual light. This compromises the usefulness of the obtained data for research in computer vision aimed at naturalistic environments.

The framework that we present here attempts to overcome this problem and achieve marker-based rigid head-pose estimation in a cluttered, naturalistic environment, without the need of attenuating visible light. Besides the advantage that the proposed method works is not bounded to near-infrared images, it is also not bounded to a specific range of orientations. Provided that the markers are placed in way that allows the non-occluded view of at least 4 markers simultaneously, pose can be estimated under any orientation.

The rest of this article is organised as follows. First, we will overview related work on maker-based motion capture in section 2. Then our proposed framework is outlined and described in section 3. In section 4 experimental results are discussed. Our overall conclusions are summarised in section 5.

2. Related work

The work that has already been done on optical motion capture is extensive. An exhaustive overview cannot be provided here, therefore. Instead, we will limit this survey to the most important work done on the specific problem of marker-based head-pose estimation.

A practical implementation of marker-based head-pose estimation has been proposed in [4]. It works with three in-

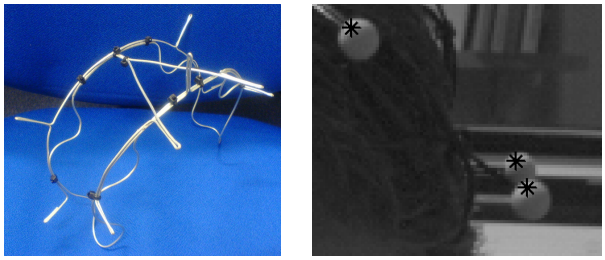
frared LED markers mounted on a set of glasses. The software tool 'FreeTrack' uses a similar method of head pose tracking, and is available online [2]. For the database presented in [1], the ground truth of head pose was estimated from 3 green LEDs placed around the face. This required having a limited intensity of the ambient illumination, as well as the absence of green colour in the background.

Apart from the method followed in [1], which works under differently restricted conditions, the limitation of currently proposed methods of optical head motion capture is that they rely on infrared-emitting markers and having the visible light attenuated by optical filtering. This means that 1) the methods will not work in environments with ambient infrared light, such as the outdoors, and 2) that a separate camera is needed for capturing ground head motion only. Another limitation is that the way the LEDs are identified in the image plane limits the freedom of rotation. The methods cannot be extended to work with 360 degree rotation. This is due to ambiguities that are inherent to solving the 2D to 3D pose inference from 3 points. However, merely adding a forth marker would not help, due to another problem that the above approaches cannot solve for rotation of 90 degrees or more. This is the problem of 'marker identification': To determine which detected marker location in the image corresponds to which marker of the target structure.

In the OPTOTRAK system (www.ndigital.com), this is solved by turning on one LED marker at a time. This means the LED markers need to be synchronised with the camera system and a series of images is required to estimate the head-pose. Unless high-speed cameras are used, this approach reduces the capture speed and causes problems with motion.

A more flexible solution to the marker-association problem is by searching for the best match between the known rigid structure and the detected marker locations. Such an approach has been adopted by Pintaric and Kaufmann [7]. Their method allows the use of multiple rigid marker-ensembles ("targets") in the same environment. The method assumes a high contrast of the markers in the infrared light spectrum and triangulation of detected marker locations between cameras. Because of the marker identification relies on triangulated point depths, multiple cameras are required. And similar most of the monocular methods mentioned above, it also depends on attenuation of visible light in order to segment the infrared-reflecting markers from the background.

The work presented here is based on the principle of finding a unique pattern of markers under a specific pose, as also followed in [7]. However, contrary to [7], the way we reduce the search space of marker identifications works with a single camera view and is robust to false positive marker detections. This means that our method is suitable for monocular pose estimation in applications where visible light needs to be captured (for other computer vision methods) and/or ambient infrared light is present (such as outdoors).



(a) (b)

Figure 1. Head-wearable marker structure. (a) The metal frame. (b) Detection of the passive white markers under a difficult lighting condition. The estimated locations of markers are indicated with '*'.

3. Motion Capture Framework

Our motion capture framework consists of several essential elements, being: (1) A light-weight structure with markers that can be easily and securely worn on the head, (2) a camera and the parameters that model its perspective- and non-linear distortion, (3) a marker-structure model that contains the relative three-dimensional locations of the markers that need to be tracked, (4) a marker-detection method that can detect and accurately localise marker-like shapes in an image with clutter, (5) an efficient perspective 3-point pose estimation procedure, (6) an initialisation procedure that assigns the detected marker locations in the image to the correct markers of the marker-structure, without prior information of pose, (7) an efficient tracking approach to limit the complexity of identifying the detected markers when the approximate pose is already known and (8) a refinement procedure to automatically adapt the marker-structure model when it might be flexed or deformed with respect to the original model.

3.1. The Light-Weight marker structure

The marker structure is formed out of 1.35mm thick galvanised metal wire and nine glued-on 'paper balls' of 10mm diameter. See figure 1. To make it black and less cold to the touch, it can be covered by black heat-shrink tubing.

The placement of markers is important, but does not require a high precision. In fact, a regularity in the placement introduces ambiguity in point correspondences. Symmetry causes an ambiguity in the direction from which the pattern is viewed. The marker structure used in our experiments has not been optimised for this purpose and has been designed to be used only with angles up to 90 degrees from frontal view. Fortunately, the imprecision of the structure's

symmetry is large enough for our pose estimation method to distinguish front from back.

But probably most importantly, the markers should not easily be occluded by each other, or by a person's head or hair.

3.2. Obtaining the Prior Models

The initial marker structure model was obtained from a close-up frontal flash-photo on a dark background, without self-occlusions, using a conventional, uncalibrated photo camera. From this, the individual markers can be easily segmented. The apparent sizes of the markers can be converted into relative distances. This very rough model is used as the seed for an iterative refinement procedure, constrained by accurately measured physical distances between pairs of the markers. If a sufficient amount of accurately measured point-to-point distances are provided, the converged result can be a highly accurate three-dimensional model.

The intrinsic calibration of the camera that will be used for pose estimation can be obtained with a flat regular checkerboard pattern and a toolbox such as Callab [8], or the Camera Calibration Toolbox for Matlab of Jean-Yves Bouguet. To accurately estimate focal length, the images of the checkerboard pattern must contain significant perspective distortion, while an accurate estimate of the non-linear distortion requires that the pattern fills the whole image.

3.3. Marker Detection and Localisaton

Figure 1 shows the markers and the detected marker locations in the images. The appearance of the passive markers depends highly on the illumination. In this example, most of the light is coming from above, slightly backwards. This changes the bright areas of the markers to a moon-like shape on the edge, and causes the detected marker locations in the image to be shifted from the middle. Because of this, the estimated 3D marker locations will be shifted by, at most, 5mm towards the light source (within the radius of the white spheres). However, such a shift does not necessarily affect the estimation of the head orientation, since all estimated marker locations are shifted similarly when the markers are lit from the same direction.

The non-distinctive and variable appearance of the markers makes it difficult to rely on local image descriptions such as SIFT or SURF features. Therefore, we have instead chosen for a generic 'bright-spot-filter', based on the principle of the Laplacian of Gaussian (LoG) filter:

$$M(\mathbf{x}, f, \tau, \gamma) = G\left\{\mathbf{x}, f, \gamma\frac{\tau}{\sqrt{3}}\right\} - \max_{k\in\{0,..,7\}} G\left\{\left(\mathbf{x} + \tau[\sin(k\frac{\pi}{4})\ \cos(k\frac{\pi}{4})]^\top\right), f, \gamma\frac{\tau}{\sqrt{3}}\right\}, \quad \tau \in \{1, 2, \ldots\}, \gamma \in, [0, 1]. \tag{1}$$

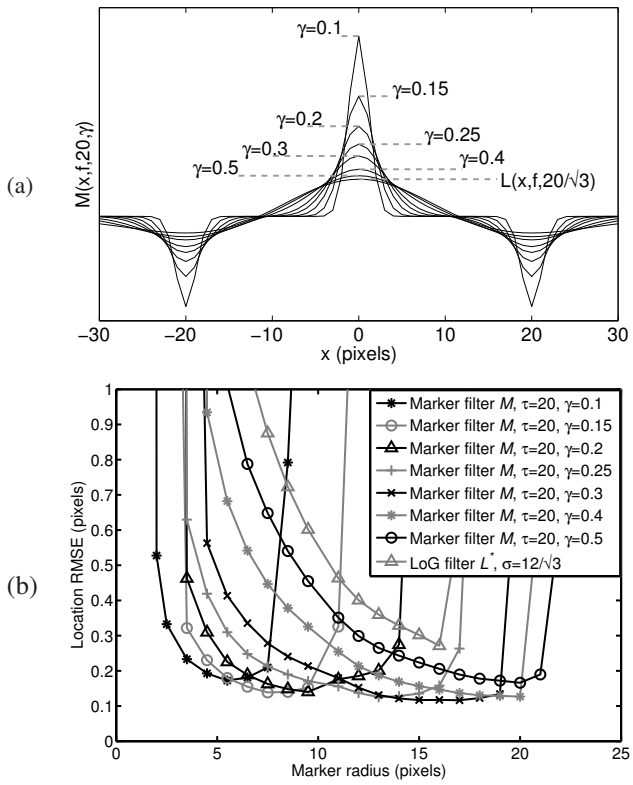


Figure 2. (a) Filter responses for different settings of γ of the adaptable marker filter $M(x, f, \tau, \gamma)$, for $\tau = 20$ and a dirac pulse 1D image function $f(x) = \delta(x)$. The response of the comparable LoG filter $L(x, f, \sigma)$ with $\sigma = \tau/\sqrt{3}$ is shown as well. (b) Root mean square error (rmse) of marker detection and localisation, measured from the two-dimensional euclidean distances between estimated locations p_s and the true marker location. Marker detection was done in synthetically generated images, of disk shapes on randomised backgrounds. Results with the proposed marker filter $M(\mathbf{x}, f, 20, \gamma)$ are shown for different values of γ. The result with a comparable LoG filter $L^*(\mathbf{x}, f, 12/\sqrt{3})$ is shown as well.

M computes the difference between the value of the Gaussian-filtered image f at position $\mathbf{x}$ and the maximum value at eight points on a circle with radius τ around $\mathbf{x}$. τ corresponds to the maximum radius of a disk-shape for which this filter is expected to work well. $G(\mathbf{x}, f, \sigma)$ is a two-dimensional Gaussian filtering of image f with standard deviation σ.

Figure 2 (a) shows the responses of $M(x, f, \tau, \gamma)$ to a dirac pulse, for different values of γ and a $\tau = 20$ pixels, compared to a LoG filter $L(x, f, \sigma)$ with scale $\sigma = \tau/\sqrt{3}$. The parameter γ allows a trade-off between invariance to the size of a disk-shaped marker with a radius smaller than τ (when $\gamma = 0$), or having less influence of clutter around a marker on the estimated location (when $\gamma \to 1$).

The optimal value of γ was determined experimentally

on synthetic images of disk shapes with random background clutter. The results of marker localisation with $M(x, f, \tau, \gamma)$ are shown in figure 2 (b) for $\tau = 20$ pixels and several values of γ. With a small γ, the accuracy for smaller markers is increased. But if γ is too small, the filter becomes sensitive to the background clutter when it is applied to larger markers. We have repeated the experiments for different values of τ, and always found the best trade-off around $\gamma = 0.3$. τ can be intuitively chosen as the maximum marker radius that can be expected in the image.

The precise location p_m of the local maximum of $M(\mathbf{x}, f, \tau, \gamma)$ is still influenced by the surrounding background clutter. To get a better estimate, the closest local maximum p_g is found in the intermediate two-dimensional Gaussian filtering $G(\mathbf{x}, f, \sigma)$. A sub-pixel correction c of the marker location is determined separately for the horizontal and vertical dimension, by first-order linear approximation, using the slopes of G on both sides of p_g:

$$c = (c_a + c_b)/2, \tag{2}$$
$$c_a = (\mu_b - \mu_a)/2d_a, \tag{3}$$
$$c_b = (\mu_b - \mu_a)/2d_b, \tag{4}$$
$$\mu_a = (g(\alpha) + g(\alpha + 1))/2, \tag{5}$$
$$\mu_b = (g(-\alpha) + g(-\alpha - 1))/2, \tag{6}$$
$$d_a = g(\alpha) - g(\alpha + 1), \tag{7}$$
$$d_b = g(-\alpha) - g(-\alpha - 1). \tag{8}$$

Here, $\alpha \in \{1, 2, \ldots\}$ is the distance at which both slopes are measured and $g(i)$ is a measure from the output of the two-dimensional Gaussian filtering at $p_g + i$, with p_g being the pixel location of the local peak in G. This estimation is based on the assumption that $g(i)$ is the shifted version of a symmetrical function $g_0(i)$, strictly increasing and differentiable for $i < 0$ and strictly decreasing and differentiable for $i > 0$. For a small horizontal shift of g_0, the difference between values at any symmetric pair of locations around 0, approximates $2\times$ the local derivative on either side. Equation 2 reduces the effect of violating this assumption, by averaging the two estimations c_a and c_b. Note that this estimation does not require concavity, nor differentiability, at the peak location of g.

To prevent erratic results when numerical derivatives d_a or d_a are close to 0, c is clipped to $[-0.5, 0.5]$. Robustness of c to such errors depends on the choice of α. We have found $\alpha = 2$ to be a good trade-off between robustness and accuracy, although the differences from using different α are small.

3.4. Efficient Perspective 3-Point Pose Estimation

Three is the minimum amount of points necessary to estimate three-dimensional pose (location and orientation) from two-dimensional image locations. Closed-form solutions can narrow the estimation down to, at most, four possible poses. However, they are known to be inefficient [6]. On the other hand, iterative algorithms require a good initial estimate.

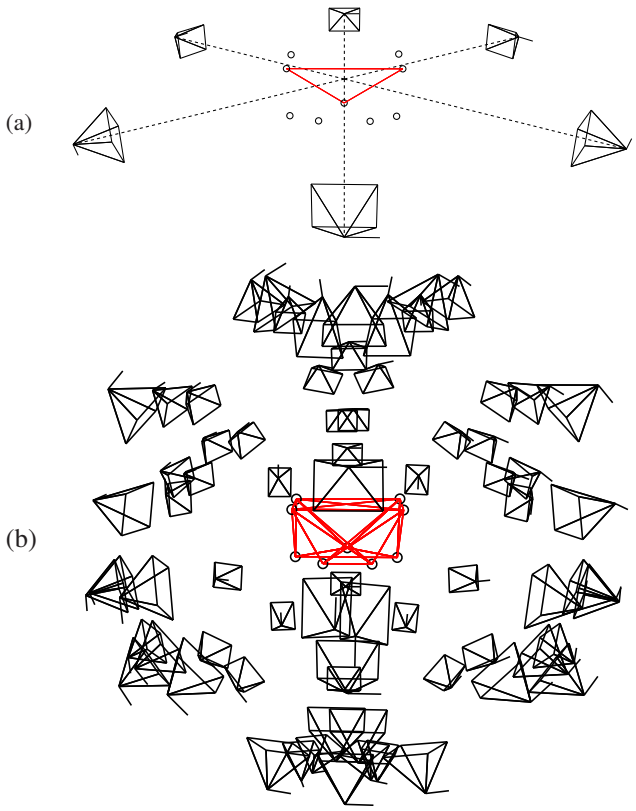


Figure 3. Reference camera orientations for triplets of points of the marker structure model. The six camera views for each triplet are centered around the mean location of the three points and perpendicular to the three different edges, from both sides in the plane spanned by the triangle. (a) shows the six views for the most equilateral triangle, (b) the 72 views for the 12 most equilateral triangles

Our approach to 3-point pose estimation is based on two important observations. First of all, out-of-camera-plane rotation cannot be estimated accurately when the three points are equidistant to the camera plane, since, from such a pose, any out-of-plane orientation change contributes to negligible change in image locations. Secondly, (near) co-linearity of the points causes pose ambiguity. This means that the computational effort can best be spent on the sets of three points that form (near) equilateral triangles with their face oriented dominantly along the camera viewing axis.

To obtain an initial pose estimate, a reference view is chosen from one of six pre-defined views for the respective triplet of points, as shown in figure 3 (a). The choice is narrowed to two views, by choosing the views perpendicular to the triplet's edge that appears the largest in the image. A

deviation from the reference pose is estimated using simple geometry that neglects perspective distortion. This initial guess is refined using the iterative Levenberg-Marquardt method.

The validity of both options is tested against one or more additional detected marker locations in the image. The locations of the nine points of the marker structure model should correspond to detected marker locations in the image, when the model is projected using the estimated pose. More details follow in section 3.5.

3.5. Pose Initialisation

Because the associations of detected marker candidates in the images with the marker structure model are not known initially, a search strategy should be applied to choose from all possible associations. Because all markers look similar in the image, they cannot be directly related to the correct points of the three-dimensional marker structure. Instead, we use a sampling consensus approach based on the 3-point method described in paragraph 3.4.

A minimal number of inliers must be set to ensure the validity of a sample, while preventing to confuse any incorrect association with detected marker locations in the image. This number can be anywhere between 4 and the number of markers in the structure. The minimum required number of non-occluded markers depends on the set range for inlier detection in the image. The more strict this can be set, the lower the number of inliers are required to prevent false point-association. The required tolerance for inlier detection depends on the accuracy of marker localisation in the image, the accuracy of the camera calibration and the accuracy of the marker-structure model. Especially before refinement of the marker-structure model, the tolerance for image locations needs to be set high, which consequently requires to raise the limit on the number of non-occluded markers.

The 84 possible combinations of 3 points out of 9 markers is reduced to a smaller set, between 10 or 20 triplets. As shown in figure 3 (b), 12 triplets already give a full coverage of the sphere of possible viewing directions and even some redundancy to handle occlusions. Still, the number of possible correspondences with image points is large. To reduce the number of image points, only moving points are considered during the initialisation step. This excludes false marker detections in a static background.

A further reduction in computation time per frame is achieved by spreading the search for triplets over multiple frames. It is better to take more frames to do the pose initialisation than to spend a long time on the correct initialisation in one frame. An initialisation will be useless if it does not represent an accurate prior for tracking in the next frame that is processed. The upper right image in figure 4 shows the first frame in which the pose has succesfully initialised (frame number 16). Frame number 15 is shown to the left.

3.6. Pose Tracking

Contrary to the pose initialisation explained in paragraph 3.5, during tracking, all of the 84 combinations of three markers are considered. Instead, the search for the correct associations of triplets is reduced in two different ways. First of all, for each of the three points in a marker triplet, only those detected marker locations are considered that are close to the back-projected locations of the estimated marker locations in the previous frame. Secondly, the estimated 3-point pose is rejected before even considering to validate it, if it is not close to the previous pose, both in location as well as orientation. This does not only reduce the possibility of false matches, but also reduces the valuable step of comparing back-projected model points to image point locations to evaluate the number of inliers.

Instead of pre-computing a rough pose estimate, the pose of the previous frame is now used as the initial pose for the Levenberg-Marquardt optimisation.

3.7. Marker Structure Refinement

Because the marker structure is flexible to allow a comfortable fit on different heads, the marker structure will be slightly different every time it is used. To ensure accurate and robust tracking, the marker-structure model has to be refined automatically. This is done by collecting poses and corresponding image locations where all markers are visible and detected in the image. The estimated poses and the original (imprecise) marker structure are then used as an initial guess for multi-view bundle adjustment that refines both the estimated poses as well as the structure. If necessary, this process can be repeated several times, or the marker structure can be updated continuously during tracking.

4. Results and Discussion

The robustness and accuracy of our proposed motion capture framework is demonstrated in figure 4. The results of head pose tracking are shown for 15 frames in a video sequence of 9.4 seconds, recorded at 60 frames per second at a resolution of 780x580 pixels. The video sequence can be viewed online at [3]. The tracking was successfully initialised in frame 16, shown in the top right (with frame 15 to its left). Although the marker structure was designed for frontal view estimation, it works for a 360 degree rotation. The current suboptimal implementation in Matlab runs at 17 frames per second on a 2GHz quad core PC. This already allows for real-time motion capture of some slower movements.

Many improvements can be added to the proposed framework to increase accuracy or robustness. First of all, the sequence of estimated poses can be filtered to eliminate jitter that comes from changes in the set of markers that are

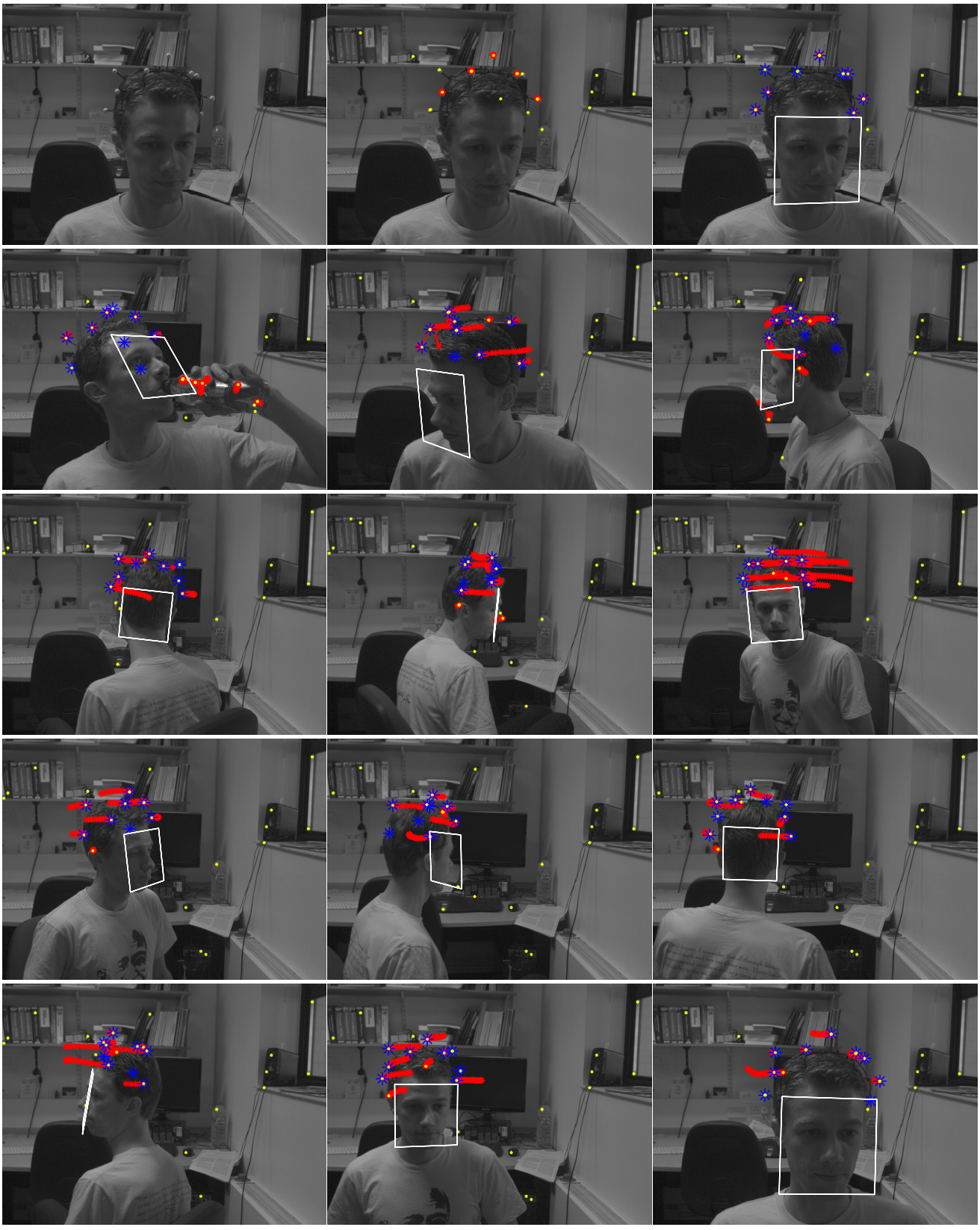


Figure 4. Frame results of head motion capture in a 9.4 seconds sequence, recorded at 60 frames per second and 780x580 pixels. Chronological order is from left to right and top to bottom. The 30 or less brightest detected marker locations are indicated with the smallest, yellow asterisks. The trails of moving markers are marked with larger, red asterisks. The 9 back-projected locations of the posed marker structure are indicated with the largest, blue asterisks. A white square indicates the estimated 3D pose of the face, relative to the marker structure and back-projected onto the image. The video sequence can be viewed online at [3].

included in the pose estimation. Secondly, an odd-one-out verification step may be added that rejects proposed marker locations with a deviating image-appearance. Thirdly, an automatic registration has to be added that accurately aligns facial points to the marker structure. When rigid facial points are detected and tracked in several frames under different head poses, their three-dimensional locations can be estimated using the estimated rigid head motion.

5. Conclusions

We have proposed a framework for monocular marker-based head-pose estimation that works under natural illumination and any orientation that leaves at least 4 markers unoccluded. The framework combines an efficient automatic pose initialisation with an efficient and robust tracking approach, as well as an automatic model-refinement procedure. We have discussed considerations and choices of how to reduce complexity or increase accuracy and robustness in each of the framework elements. Furthermore, we have demonstrated successful tracking under fast 360 degree head motion.

Acknowledgment

This work was supported by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB).

References

[1] S. Asteriadis, D. Soufleros, K. Karpouzis, and S. Kollias. A natural head pose and eye gaze dataset. In *AFFINE '09 Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots*, 2009.

[2] Babasior. *FREETRACK Handbook V2.1*. Free Track, October 2007.

[3] J. F. Lichtenauer and M. Pantic. Head motion capture. Intelligent Behaviour Understanding Group, Imperial College London, http://ibug.doc.ic.ac.uk/research/head-motion-capture/, 2011.

[4] S. Meers, K. Ward, and I. Piper. *Mechatronics and Machine Vision in Practice*, chapter Simple, Robust and Accurate Head-Pose Tracking Using a Single Camera, pages 111–122. Springer Berlin Heidelberg, December 2007.

[5] T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis; special issue on modeling people: Vision-based understanding of a person's shape, appearance, movement and behaviour. *Computer Vision and Image Understanding*, 104(2-3):90–126, 2006.

[6] F. Moreno-Noguer and P. Fua. Accurate non-iterative o(n) solution to the pnp problem. In *IEEE International Conference on Computer Vision*, Rio de Janeiro, Brazil, October 2007. IEEE.

[7] T. Pintaric and H. Kaufmann. Affordable infrared-optical pose-tracking for virtual and augmented reality. In *Proceedings of Trends and Issues in Tracking for Virtual Environments Workshop, IEEE VR*, 2007.

[8] K. H. Strobl, W. Sepp, and S. Fuchs. Callab 2005 and calde. Inst. of Robotics and Mechatronics, German Aerospace Center, http://www.robotic.dlr.de/callab/, 2005.

[9] H. Zhou and H. Hu. Human motion tracking for rehabilitation–a survey. *Biomedical Signal Processing and Control*, 3(1):1–18, 2008.