A Multi-Sensor Visual Tracking System for Behavior Monitoring of At-Risk Children

Ravishankar Sivalingam[*] Anoop Cherian[*] Joshua Fasching[*] Nicholas Walczak[*] Nathaniel Bird[#]
Vassilios Morellas[*] Barbara Murphy[§] Kathryn Cullen[$] Kelvin Lim[$]
Guillermo Sapiro[†] Nikolaos Papanikolopoulos[*]
[*]{ravi, cherian, fasching, walczak, morellas, npapas}@cs.umn.edu
{[†]guille, [$]rega0026, [$]kolim, [§]murph028}@umn.edu [#]n-bird@onu.edu
[*]Dept. of Computer Science & Engg. [†] Dept. of Electrical & Computer Engg.
[§]Institute of Child Development [$]Center for Magnetic Resonance Research, Medical School
[*†$]University of Minnesota, MN 55455
[#]Dept. of Electrical & Computer Engg. & Computer Science, Ohio Northern University, OH 45810

***Abstract*— Clinical studies confirm that mental illnesses such as autism, Obsessive Compulsive Disorder (OCD), etc. show behavioral abnormalities even at very young ages; the early diagnosis of which can help steer effective treatments. Most often, the behavior of such at-risk children deviate in very subtle ways from that of a normal child; correct diagnosis of which requires prolonged and continuous monitoring of their activities by a clinician, which is a difficult and time intensive task. As a result, the development of automation tools for assisting in such monitoring activities will be an important step towards effective utilization of the diagnostic resources. In this paper, we approach the problem from a computer vision standpoint, and propose a novel system for the automatic monitoring of the behavior of children in their natural environment through the deployment of multiple non-invasive sensors (cameras and depth sensors). We provide details of our system, together with algorithms for the robust tracking of the activities of the children. Our experiments, conducted in the Shirley G. Moore Laboratory School, demonstrate the effectiveness of our methodology.**

I. INTRODUCTION

Recent studies [1] have shown that the prevalence of diseases in the category of Autism Spectrum Disorders (ASDs) are in the range of 6.5 per 1000. These diseases lead to neuro-motor dysfunctions causing disabilities for the patient to gather social and communication skills. Fortunately, it has been shown that early diagnosis [2] (within less than 1-2 years of age) of such diseases can help the patient gain additional skills to overcome the difficulties or in some cases even completely recover [3]. The diagnosis of such mental disorders requires the consistent monitoring for abnormalities in the activities of the children. Sometimes these deviations are so subtle (for example: abnormal motion gestures such as walking on toes, sitting clumsily for extended periods, abnormal facial expressions, etc.) that it requires a qualified doctor to analyze video footages of the at-risk children for hours [4], [5]. Considering the scarcity of such specialized medical practitioners compared to the volume of patients, along with considering the high-cost involved, there is a strong motivation for investigating technological alternatives to solve the problem.

Fig. 1. A pre-kindergarten classroom setting used for collecting the video data for our experiments.

In this paper, we present a computer vision system utilizing cameras and depth sensors to track children across multiple views effectively and for extended periods of time. The goal of this system is to monitor the children in their natural environment and detect risk markers for developmental disorders. Although computer vision has been used previously for patient monitoring, such as human gait analysis [6], facial expression analysis [7], etc., monitoring of the activities of children is an extremely challenging activity as their motion is less structured compared to adults'. Previous work in such monitoring scenarios used either surgically implanted fiducial markers [8], or infra-red markers [9]. Such intrusive mechanisms might be unappealing for our application as: (i) the behavior of the children might be severely influenced by these markers, and (ii) they might not be scalable towards effective whole body tracking including facial expressions. Towards this end, we propose to use multiple cameras for the monitoring activity, combined with 3D depth sensors. In this work, we primarily use the recently introduced Microsoft Kinect sensor [10] due to its low cost and ease of deployment. To the best of our knowledge, it is the first time that such a system has been envisaged for this application. Figure 1 shows a snapshot from a real medical lab school, the data from which are used in our experiments.

The paper is organized as follows. In Section II, we start with a brief review of previous literature in people tracking to set the stage for this paper. In Section III, we state the

goal of this paper, followed by a brief overview of the sensors that we use and an analysis of their operating characteristics. In Section IV, we describe our algorithms for clustering, robust tracking and track alignment, which precedes the experimental results and conclusion in Sections VI and VII, respectively.

II. Related Work

Object tracking is one of the most researched topics in the computer vision community with many effective solutions already in large scale deployment. A survey of some of the recent advances in this area can be found in [11]. Inferring the 3D structure of the scene and the objects within is an important requirement for effective video analytics, such as knowing the object size, segmenting an object from clutter, etc. Various configuration modalities of such multi-view camera tracking systems can be found in [12], [13]. Unfortunately, such systems require at least two calibrated cameras looking at the object of interest, which might not always be possible due to occlusion, or scene clutter, especially in a classroom environment where there might be limited alternatives for mounting the cameras. In addition, reconstructing a classroom full of children needs a thorough understanding of the scene dynamics. A system that is closer in spirit to our approach is by Kanade et al. [14], which could achieve 3D scene reconstructions at 15 frames per second. On the downside, this setup requires severe modifications of the environment which is not viable considering our final goal of natural behavior analysis.

Looking at the problem from another perspective, a combination of cameras with depth sensors is an alternative to achieve the same result. Methods presented in papers such as [10], [15] propose various applications of this combination for a variety of tasks using the Kinect platform. A major deterrent in using multiple such sensors together is the need to deduce a global frame of reference through their combined calibration. As the depth sensors are often found to be very noisy, such calibration is not always very accurate, which is an important problem that we address in this paper.

An effective representation of the objects being tracked in the scene is an important ingredient of any tracking algorithm. Although there are a variety of choices such as mean-shift tracker [16], or SIFT based tracker [17], we resort to region covariance descriptors introduced in [18] for the task at hand due to their robustness to camera noise, scene illumination, partial occlusions, and affine transformations. Moreover, these descriptors are efficient with respect to storage and can be computed cheaply from the basic geometric features of the tracked objects. We provide novel extensions of such covariance based tracking for our specific scenario through combining the descriptor with a dictionary based approach for fast object recognition, together with a Kalman filter for improving the robustness of recognition.

III. Problem Description

As a first step towards achieving the goal of a comprehensive system for monitoring the mental health of at-risk children, in this paper, we primarily discuss the problem of multi-sensor object tracking in an unstructured classroom environment. As recording of even the most subtle details regarding each child is of primary importance while analyzing for abnormal behaviors, the proposed system has to deal with resolving occlusions such that the children are observable in at least a subset of the deployed sensors. Using multiple sensors in multiple modalities naturally brings along the problem of their calibration with respect to a global frame of reference. As the sensors (especially the Kinect depth sensors) are found to be limited in accuracy with respect to their perceivable depth, we propose ideas for their combined calibration.

A second problem that we deal with in this paper is the tracking of objects across sensors; or the more challenging problem of labeling objects that go out of the scene perimeters and re-enter. we propose the use of covariance based descriptors on 3D point clouds in combination with a dictionary learning framework for modeling the appearances of the objects in the scene and for fast matching. Due to errors in the sensor measurements, it is often seen that there is usually a minor misalignment of the inferred object tracks by the sensors, for which we provide a Kalman filter based realignment solution.

A. Proposed System

Compared to the popular image+depth motion sensors such as the Vicon Multi-camera system [19] or the time-of-flight range sensors such as the SwissRanger systems [20], Microsoft Kinect provides an inexpensive and portable monitoring platform, which along with providing depth information, also provides good quality color image data. In addition, the system is packaged with software interfaces for a preliminary behavior analysis system that proves useful in our proposed system. The device equips a near infrared light system for the depth estimation that is insensitive to changes in illumination; a common problem found in indoor settings. As a result, we decided to use multiple Kinects in our classroom setting. Before going into the details of calibrating these RGB+D sensors, it is worth looking at some of the operating characteristics of the sensor we use.

B. Measurement Quality

To quantify the depth measurement accuracy of the Kinect sensor for the particular task at hand, the following experiment was conducted. A Kinect, mounted on a tripod was placed in front of a wall; its distance to the wall changed gradually, simultaneously recording the distance to the wall as provided by the depth sensor. Assuming K is the matrix of intrinsic camera parameters, x_{ij} is the homogeneous image coordinate ($x_{ij} = (i, j, 1)^T$), and d_{ij} represents the depth measured at the (i, j)th image coordinate, then we have the following relation between the location of the actual world point X_w and the sensed point x_{ij}:

$$X_w = d_{ij} K^{-1} x_{ij} \tag{1}$$

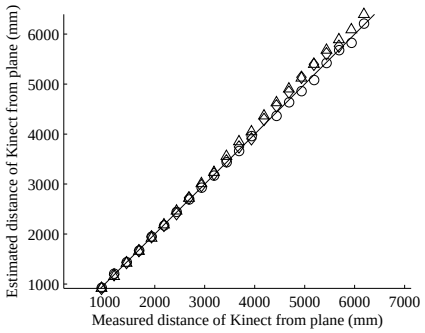


Fig. 2. Kinect depth measurement error. The straight line corresponds to the ground truth distance to the wall, and the different symbols represent the average distance estimated by different Kinect sensors. This plot demonstrates that the estimation error in the depth sensor increases with the distance from the camera, and that different Kinects have slightly different characteristics.

As we have a wall in front of the camera, all these world points should lie on the same plane. The parameters of this plane is computed through linear least squares on (1) for many such camera coordinates. Figure 2 shows a plot of the estimated distance against the actual distance (measured manually). It is clear from the plot, the quality of the depth measurement seems reasonably accurate within a range of 1–3 meters, while beyond that the noise seems to increase gradually.

This analysis was taken into account while deciding on the placement of the sensors in our classroom setting. Our system deployed four RGB+D sensors around the classroom, each positioned in such a way that each sensor provided a novel view of the room; the views of the sensors were not completely disjoint, so that we can account for noise through a realignment mechanism that we will discuss in a later part of the paper.

C. Multi-Kinect Calibration

A good 3D reconstruction is inevitably dependent on good calibration of all the cameras and their sensors. We approach the problem systematically; first calibrating the cameras with respect to their depth sensors, and then calibrating all the depth sensors against each other. For the former task, a set of world points whose known real world coordinates are chosen, followed by computing the intrinsic camera calibration matrix through the standard Direct Linear Transform (DLT) algorithm [21].

Calibration of the sensors with respect to each other was done using a rigid calibration rig made of PVC piping as large as 1.5m, marked at regular intervals (Figure. 3). This setup not only helps in spanning the field of view of multiple sensors, but also minimizes self occlusion, aiding the use of majority of the marked points on it. Next, the projection matrices for each camera with respect to the world coordinates were computed using these marked rig points using the Gold Standard algorithm [21].

IV. Segmentation & Representation

In this section, we describe algorithms for segmenting out each subject from the 3D point clouds and representing each object by an appearance descriptor. These segments are then used for tracking and higher level behavioral inference.

A. Euclidean Point Clustering

At each instant t from each camera c, we obtain a 3-dimensional point cloud in a global reference frame, from the corresponding depth image and the camera calibration. We then segment this point cloud into separate objects through agglomerative hierarchical clustering based on the Euclidean distances. Based on the tolerance, we get distinct segments corresponding to the objects in the scene. At this point, no distinction has been made between the background and foreground objects. With the help of efficient background subtraction techniques such as [22], [23], [24], we may be able to speed up this process by only considering the points in the cloud corresponding to foreground objects of interest, as well as prevent the merging of any subject with a background object he/she may be in contact with.

Once the point cloud has been segmented, the labels and the bounding boxes for each object are then reprojected onto the depth image. These labels are used to identify the RGB pixel information relevant to each 3D point in a cluster.

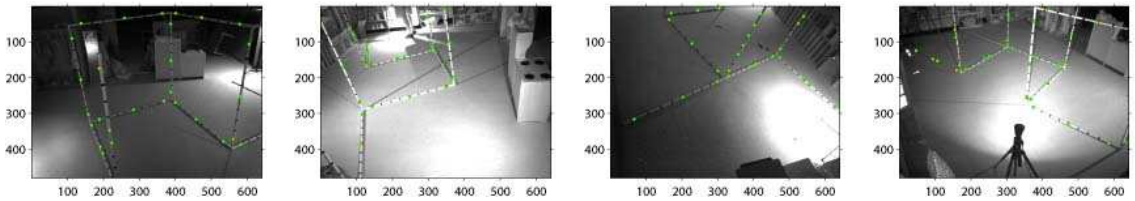
Although each camera has been calibrated with reference to a global frame, due to different fields-of-view and to minimize any effect of inaccurate camera calibration, this segmentation process is done independently on each camera's input. After this step, we now have a colored point cloud segmented into individual objects, for each camera.

B. Covariance Descriptors on 3D Point Clouds

In the previous step, we had segmented each 3-dimensional point cloud into separate objects. In this section, we discuss a compact but powerful representation for each object in the point cloud, which will not only assist us in tracking within a single camera view, but also facilitates estimating fast correspondences between objects in different camera views. For the i^{th} segment $S_{(i,t,c)}$ in frame t and camera c, we compute the region covariance descriptor $C_{(i,t,c)}$ [25] over features derived from the color and depth information. As the name implies, the $n \times n$ feature descriptor is computed as the covariance of the n-dimensional feature vector at each pixel in the region of interest.

In this paper, we use an $n = 12$-dimensional feature vector consisting of (x,y) image coordinates, *RGB* color information, first and second derivatives of the image intensity I_x, I_y, I_{xx}, I_{yy}, I_{xy}, the gradient magnitude $\sqrt{I_x^2 + I_y^2}$, and the gradient orientation $\theta = \arctan\frac{I_y}{I_x}$. This extensive descriptor combines information from the sensors in an efficient way, and gives us a fast way of establishing correspondences across multiple cameras and also helps in re-identifying individuals who return to the scene after temporary absence. The depth value and the derivatives of the depth map can also be easily integrated with the image features in the covariance descriptor, but for the tracking experiments they did not offer any improved performance.

Since these descriptors lie on a Riemannian manifold and not a Euclidean space, their pair-wise distances are computed



(a) Camera 1 (b) Camera 2 (c) Camera 3 (d) Camera 4

Fig. 3. The rig setup used to calibrate the cameras and the depth sensors.

using a geodesic distance metric involving computation of generalized eigenvalues between all pairs of descriptors, leading to a high computational cost. Instead, we use the Log-Euclidean distance metric between covariances [26] given by:

$$D_{\mathrm{LE}}(C_1, C_2) = \|\log C_1 - \log C_2\|_F, \tag{2}$$

where log refers to the matrix logarithm. This involves computing the eigen-decomposition of individual matrices only and is thus far more efficient. The matrix logarithm of the positive definite covariance matrices produce symmetric matrices that form a Euclidean space, and the usual Euclidean distance is computed in this space. Due to the symmetry of the matrices, it is sufficient to retain only the upper triangular part of the log-covariances, yielding an $n(n+1)/2$-dimensional Euclidean feature vector. However, in order to account for the remaining values, we scale the off-diagonal elements by $\sqrt{2}$ during vectorization.

Although these descriptors may not be sufficient to recognize our tracking subjects across multiple sessions spanning over days (since their appearances might change due to changes in attire), recall that our application is mainly assistive and we assume some level of semi-supervision from the medical personnel in resolving such tracking difficulties. If there is a specific subject that is flagged as having salient behavioral patterns, the doctor can interactively mark that person in each session. The daily tracks are then consolidated using this information and a long-term activity analysis can be performed.

In our current setting, we are using the 640×480 resolution Kinect RGB camera which is insufficient to recognize faces effectively at these distances. To the best of our knowledge, there are no existing computer vision algorithms that can recognize a person's appearance at this resolution across multiple days with changes in clothing.

V. Tracking via Covariance Descriptors

Tracking the motion of the subjects is an important step in behavior analysis. For example, one of the behavioral cues that medical personnel will be interested in looking at in a diagnostic session will be the hyperactivity of children. Thus robust motion tracking is essential. As discussed in the previous section, we now have segmented out three-dimensional objects from multiple cameras in a global frame of reference, and region covariance descriptors for each segment computed. In this section, we propose mechanisms to track each of the subjects using Kalman filters; the state vector consisting of the position $[x,y,z]^T$ of the subject derived from the centroid of their point cloud, its velocity $[v_x, v_y, v_z]^T$ and the half-vectorized log-Euclidean covariance descriptor C that represents its appearance and shape.

A. Robust Tracking

We used a codebook based approach to reliably track the subjects via their covariance descriptors. The vectorized log-covariance descriptors from the multiple sensors are averaged to produce a mean descriptor that is kept as one of the codebook atoms; the label of this atom being set to that of the subject being tracked. Given a new frame, a one-against-all matching is performed with the covariance descriptor from the region in the new frame of the subject against all the atoms in the codebook; the atom with the minimum Euclidean distance returning a successful match. The respective atom of the codebook is then updated with this new descriptor by computing their weighted mean.

The codebook aids in resolving three issues: (i) clustering the appearances from the multiple sensors into a global appearance repository and updating each appearance with its latest covariance descriptor, (ii) keeping uncertainty regarding the clusters (through computing a covariance of the vectorized descriptors, that helps in deriving a Mahalanobis distance type extension to the Log-Euclidean distance metric for more robust matching), and (iii) book-keeping information for subjects that go out of the bounds of the fields of view and reappear later. Temporal information is also maintained for each cluster to prune out those subjects that have not been tracked for a long time, or for which there has not been any significant motion detected. Static objects in the area (produced due to incorrect segmentation) are easily removed through this approach as they will have very small cluster sizes as the changes to their appearances are meager. The dictionary size is kept constant in our design for computational reasons.

Since the parts of the same object seen by different camera views are different, the point cloud from each camera returns only a portion of the entire 3-dimensional shape of the object. Therefore any centroid computed from this partial point cloud will be offset from the true object centroid, biased towards the location of the camera. Hence the tracked centroid positions from each camera must be combined to obtain the true object centroid. This is performed for each object by averaging its observed centroids over all the relevant views.

VI. Experiments and Results

In this section, we detail the experiments that we performed to substantiate the effectiveness of our system. To-

wards this end, we decided to test the three of the major components of our methodology: (i) usage of covariance matrices for tracking and calibration of the subjects in their natural environment, (ii) the robustness to track alignment provided by the Kalman filter, and (iii) an application of the system to a preliminary analysis of the behaviors of our subjects. Each of these analyses was conducted on multiple videos captured from a pre-kindergarten classroom lab school, the Shirley G. Moore Laboratory School, across various sessions. Each session involved children in the age of 3–5 and consisted of approximately 10 kids. The kids were allowed to freely move around in the classroom and their videos captured using four Kinect sensors. Approval for this study was obtained from the Institutional Review Board at the University of Minnesota. In this section we show results from two of these videos (referred to here as Scene A and Scene B) with 4 cameras. However the interested reader is referred to our website[1] for up-to-date videos and results.

A. Segmentation and Recognition

As described in the previous sections, we used a set of 12 features to create 12×12 covariance matrices for each subject from its 3D appearance to track the subjects from camera to camera, creating their trajectory profiles. One of the important cues that the medical practitioners are interested in is to understand the hyperactivity (for *e.g.*, very active children, running around, leader-follower movements, etc.) of the subjects. Later with the tracking information we will demonstrate the estimation of activity levels of the children observed in the scene.

B. Tracking and Track Alignment

The covariance descriptors are used for frame-to-frame correspondence estimation in the tracking as well as for correspondence across multiple cameras. Figure 4 shows the tracking across multiple frames in the same camera view. The children and the adults are tracked well even during partial occlusions and temporary exit from the scene. When the tracked objects are not seen for more than a threshold τ frames, they are considered lost and new trackers are initialized. Therefore, during full occlusions greater than τ frames, the tracker loses the individual it is tracking and new trackers are initialized when they are seen again.

The correspondence between objects from multiple cameras are estimated based on their covariance descriptors, and the tracks from the different cameras are merged. The true centroid of the object is estimated from the mean of the observed centroids from the individual cameras. Figure 5 shows this process for a single track from Scene B. The partially observed point-clouds centroids along the track and the estimated object centroid are shown, with correspondences marked by connecting lines.

C. Behavioral Analysis

Although the possibilities of behavioral analysis with our system are tremendous, we detail below one application that

[1]http://baarc.cs.umn.edu/

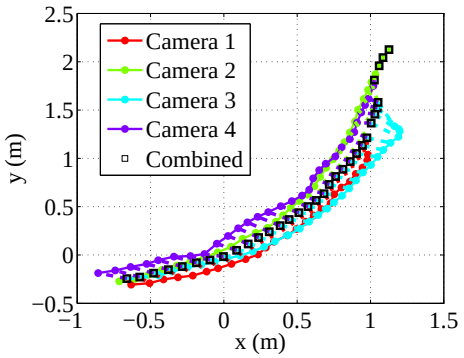


Fig. 5. Object centroids viewed from multiple cameras and the merged true centroid. The partial point clouds viewed from the different cameras yield slightly offset centroid locations. The object true centroid at time t is estimated as the mean of the corresponding observed centroids.

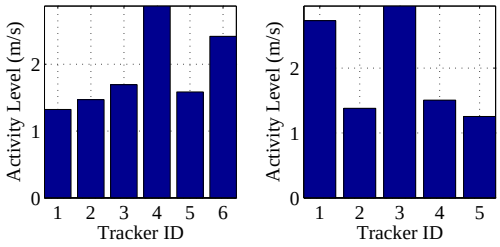


Fig. 6. Plot showing the activity levels of the tracked objects in Scene A (left) and Scene B (right). There is no classification into humans and background yet, and so some of the background objects also show a very insignificant amount of activity due to sensor noise.

can be very easily achieved using our setup, and which we found is one of the significant pointers towards a behavioral abnormality; *motion tracks* of the subjects. To this end, we segregated the classroom into a virtual 3D accumulator grid, as viewed by the multiple sensors, with the goal of computing the frequency with which subjects occupy each accumulator cell. Motion tracks of the subjects using the covariances, superimposed on this grid provides an easy check on how active the children were and how they were behaving in the scene. The average activity level of the child is also estimated from the Kalman trackers and the instantaneous trajectory coordinates. The activity levels for Scene A and Scene B are shown in Figure 6. This shows a preliminary insight into the kind of capabilities possible with this system.

Another possibility is to study the velocity profiles of the subjects, and identify salient profiles. For *e.g.*, children who are not very active can be recognized as interesting, based on the saliency of their velocity information compared to that of most other children. Periodicity in actions can point to OCD-type actions, such as passing multiple times through a doorway, or repeatedly touching certain objects. These behaviors can be identified by finer tracking of different parts of the subjects such as hands and feet.

VII. Conclusion

In this paper, we proposed a novel application of a multi-sensor setup for monitoring children at very young ages towards the goal of finding serious behavioral abnormalities. We provided details of our system setup using the

Fig. 4. Tracking within a single camera - Scene A Camera 1 frames 40, 80, 120, 160, 200, and 240.

Microsoft Kinect platform along with techniques to calibrate the devices. A covariance based descriptor on the point clouds was used to track the subjects in the scene effectively across frames and sessions, followed by a Kalman filter based approach to tackle noise in the depth sensors for robust tracking. Although covariance descriptors are useful for holistic motion tracking, some of the behavioral analysis requires more fine grained information such as the motion of the limbs of the children. Thus one direction to build upon the current system is to use dense optical flow of the point clouds. Another important issue that was not considered in this paper is to detect stationary subjects. Also, as mentioned earlier the appearance models we are currently using cannot be used to recognize individuals across multiple days when their clothing changes. However, with strategically placed high-resolution cameras, we may employ face recognition techniques to identify the same individual across multiple days, linking their daily appearance models. We plan to tackle these issues in the future.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation through grants #IIP-0443945, #CNS-0821474, #IIP-0934327, #CNS-1039741, and #SMA-1028076. Additional support for GS from NGA, ONR, ARO, and NSSEFF (AFOSR) is also gratefully acknowledged.

REFERENCES

[1] S. Myers, C. Johnson, *et al.*, "Management of children with autism spectrum disorders," *AAP Policy*, vol. 120, no. 5, p. 1162, 2007.

[2] E. Walker, T. Savoie, and D. Davis, "Neuromotor precursors of schizophrenia," *Schizophrenia Bulletin*, vol. 20, no. 3, pp. 441–451, 1994.

[3] M. Helt, E. Kelley, M. Kinsbourne, J. Pandey, H. Boorstein, M. Herbert, and D. Fein, "Can children with autism recover? if so, how?," *Neuropsychology review*, vol. 18, no. 4, pp. 339–366, 2008.

[4] H. Massie, "Blind ratings of mother-infant interaction in home movies of prepsychotic and normal infants," *American Journal of Psychiatry*, vol. 135, no. 11, p. 1371, 1978.

[5] E. Walker, K. Grimes, D. Davis, and A. Smith, "Childhood precursors of schizophrenia: facial expressions of emotion," *American Journal of Psychiatry*, vol. 150, no. 11, p. 1654, 1993.

[6] E. Ribnick, V. Morellas, and N. Papanikolopoulos, "Human motion patterns from single camera cues for medical applications," in *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pp. 5102–5107, IEEE.

[7] C. Shan, S. Gong, and P. McOwan, "A comprehensive empirical study on linear subspace methods for facial expression analysis," in *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW '06. Conference on*, p. 153, june 2006.

[8] R. MacKay, P. Graham, J. Logue, and C. Moore, "Patient positioning using detailed three-dimensional surface data for patients undergoing conformal radiation therapy for carcinoma of the prostate: a feasibility study," *International Journal of Radiation Oncology* Biology* Physics*, vol. 49, no. 1, pp. 225–230, 2001.

[9] N. Linthout, D. Verellen, K. Tournel, and G. Storme, "Six dimensional analysis with daily stereoscopic x-ray imaging of intrafraction patient motion in head and neck treatments using five points fixation masks," *Medical physics*, vol. 33, p. 504, 2006.

[10] 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," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR11), Colorado Springs, USA*, 2011.

[11] T. Moeslund, A. Hilton, and V. Kruger, "A survey of advances in vision-based human motion capture and analysis," *Computer vision and image understanding*, vol. 104, no. 2-3, pp. 90–126, 2006.

[12] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera people tracking with a probabilistic occupancy map," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 267–282, 2008.

[13] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1, pp. 519 – 528, june 2006.

[14] T. Kanade, P. Rander, and P. Narayanan, "Virtualized reality: constructing virtual worlds from real scenes," *Multimedia, IEEE*, vol. 4, pp. 34 –47, jan-mar 1997.

[15] I. Oikonomidis, N. Kyriazis, and A. Argyros, "Efficient model-based 3d tracking of hand articulations using kinect," *BMVC, Aug*, vol. 2, 2011.

[16] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on pattern analysis and machine intelligence*, pp. 603–619, 2002.

[17] H. Zhou, Y. Yuan, and C. Shi, "Object tracking using sift features and mean shift," *Computer Vision and Image Understanding*, vol. 113, no. 3, pp. 345–352, 2009.

[18] F. Porikli, O. Tuzel, and P. Meer, "Covariance tracking using model update based on lie algebra," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1, pp. 728–735, IEEE, 2006.

[19] Vicon, *Vicon MX Systems*, October 2006.

[20] Mesa Imaging, *SR4000 SwissRanger Camera*.

[21] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge Univ Press, 2000.

[22] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1999.*, vol. 2, pp. 2 vol. (xxiii+637+663), 1999.

[23] K. Patwardhan, G. Sapiro, and V. Morellas, "Robust foreground detection in video using pixel layers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 746 –751, april 2008.

[24] R. Sivalingam, A. D'Souza, V. Morellas, N. Papanikolopoulos, M. Bazakos, and R. Miezianko, "Dictionary learning for robust background modeling," in *IEEE International Conference on Robotics and Automation (ICRA), 2011*, pp. 4234 –4239, may 2011.

[25] O. Tuzel, F. Porikli, and P. Meer, "Region covariance: A fast descriptor for detection and classification," in *Computer Vision ECCV 2006* (A. Leonardis, H. Bischof, and A. Pinz, eds.), vol. 3952 of *Lecture Notes in Computer Science*, pp. 589–600, Springer Berlin / Heidelberg, 2006.

[26] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Log-Euclidean metrics for fast and simple calculus on diffusion tensors," *Magnetic Resonance in Medicine*, vol. 56, pp. 411–421, August 2006. PMID: 16788917.