ISSN XXXX XXXX © 2017 IJESC

Research Article Volume 7 Issue No.8

A Survey Utility Person Detection Multi-View Video Tracking Annotation Model

Dr. P.Sumitra[1], M.Senbagapriya, M.Phil [2]
Assistant Professor[1], Research Scholar[2]
PG and Research Department of Computer Science
Vivekananda College of Arts and Sciences for Women (Autonomous), Elayamapalayam, India

Abstract:
In this thesis a generic methodology for the semi-automatic generation of reliable position annotations for evaluating multi-camera people-trackers on large video data sets. Most of the annotation data are automatically computed, by estimating a consensus tracking result from multiple existing trackers and people detectors and classifying it as either reliable or not. A small subset of the data, composed of tracks with insufficient reliability, is verified by a human using a simple binary decision task, a process faster than marking the correct person position. The proposed framework is generic and can handle additional trackers. In this thesis studied the most commonly use face edge detection techniques of Enhnaced Sobel Edge Annotation Algorithm (ESEAA). Higher-level edge detection techniques and appropriate programming tools only facilitate the process but do not make it a simple task.

KEYWORDS: Image processing, Digital Image Processing, Analog Image Processing Two dimensional signals

1. INTRODUCTION

Image processing is processing of images using mathematical operations by using any form of signal processing for which the input is an image, such as a photograph or video frame; the output of image processing may be either an image or a set of characteristics or parameters related to the image. Most image-processing techniques involve treating the image as a two-dimensional signal and applying standard signal-processing techniques to it. Image processing usually refers to digital image processing, but optical and analog image processing also are possible. Digital image processing is the use of computer algorithms to perform image processing on digital images. As a subcategory or field of digital signal processing, digital image processing has many advantages over analog image processing. It allows a much wider range of algorithms to be applied to the input data and can avoid problems such as the build-up of noise and signal distortion during processing. Since images are defined over two dimensions (perhaps more) digital image processing may be modeled in the form of multidimensional systems.

Purpose of Image processing

The purpose of image processing is divided into 5 groups. They are:

1. Visualization - Observe the objects that are not visible.
2. Image sharpening and restoration - To create a better image.
3. Image retrieval - Seek for the image of interest.
4. Measurement of pattern – Measures various objects in an image.
5. Image Recognition – Distinguish the objects in an image.

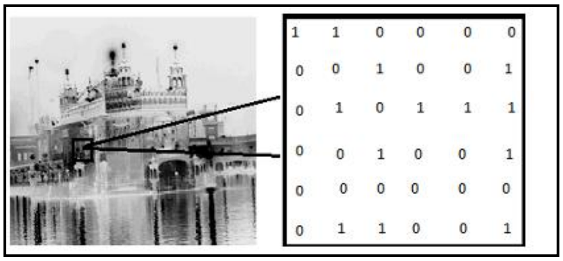
Digital Processing techniques help in manipulation of the digital images by using computers. As raw data from imaging sensors from satellite platform contains deficiencies. To get over such flaws and to get originality of information, it has to undergo various phases of processing. The three general phases that all types of data have to undergo while using digital technique is Pre- processing, enhancement and display, information extraction.

1.1.1 IMAGE AND ITS TYPES

An image may be well-defined such as a two-dimensional function F (a, b).Where a and b are spatial (plane) coordinate, and the amplitude of F at any pair of coordinates (a, b) is called the intensity or gray level of the image at that point. When a, b and the amplitude values of are all predetermined discrete quantity, we will call the image as digital image. A digital image is collection of a finite number of elements, in which each element has a certain value and location. These elements of digital image are known as image elements, picture elements, pels, and pixels. Pixel is the word mostly used refers to the elements of a digital image.

1.1.2 TYPES OF DIGITAL IMAGES:

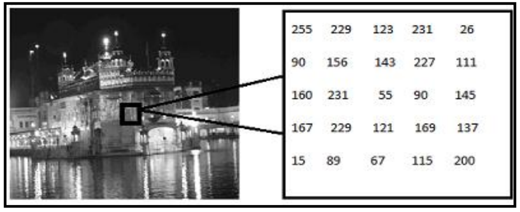
Binary: In binary image the value of each pixel is either black or white. The image have only two possible values for each pixel either 0 or 1, we need one bit per pixel.



1	1	0	0	0	0
0	0	1	0	0	1
0	1	0	1	1	1
0	0	1	0	0	1
0	0	0	0	0	0
0	1	1	0	0	1

Figure.1. Binary Images

Grayscale: In grayscale image each pixel is shade of gray, which have value normally 0 [black] to 255 [white]. This means that each pixel in this image can be shown by eight bits, hat is exactly of one byte. Other grayscale ranges can be used, but usually they are also power of 2.



255	229	123	231	26
90	156	143	227	111
160	231	55	90	145
167	229	121	169	137
15	89	67	115	200

Figure.2. Grayscale Images

True Color or RGB: Each pixel in the RGB image has a particular color; that color in the image is described by the quantity of red, green and blue value in image. If each of the components has a range from 0–255, this means that this gives a total of 2563 different possible colors values. That means such an image is "stack" of three matrices; that represent the red, green and blue values in the image for each pixel. This way we can say that for every pixel in the RGB image there are corresponding 3 values. Indexed: Mostly all the colors images have a subset of more than sixteen million possible colors. For ease of storage and handling of file, the image has an related color map, or we can say the colors palette, that is simply a list of all the colors which can be used in that image. Each pixel has a value associated with it but it does not give its color as for as we see in an RGB image. Instead it gives an index to the color in map. It is convenient for an image if it has 256 colors or less. The index values will require only one byte to store each. Some image file formats such as GIF which allow 256 colors only.

1.3.3 DIGITAL IMAGE FILE TYPES

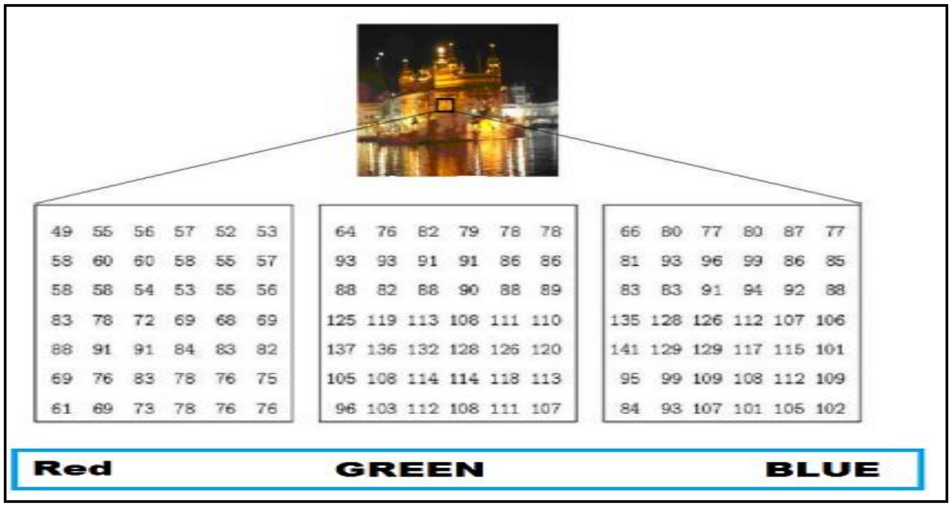
BMP:

Bmp stands for Bitmap. Every picture on a computer appears to be a BMP. In Windows XP the Paint program save its images automatically in bitmap format, however in Windows Vista images are saved now into JPEG format. Bitmap is the basis platform for many other file types. Benefits: High quality image, Easy to change and edit, No loss in image through process

Downfalls: Difficulty while displayable on internet and large in file size. Digital camera manufacturers obviously see the value in high quality images that eventually take up less space.

Benefits: Small size image, easily viewable from internet, Use millions of colors, and perfect for many type of images

Downfalls: High compression loses quality of image, every time a JPG image is saved; it loses more and more quality of the picture.



49	55	56	57	52	53
58	60	60	58	55	57
58	58	54	53	55	56
83	78	72	69	68	69
88	91	91	84	83	82
69	76	83	78	76	75
61	69	73	78	76	76

64	76	82	79	78	78
93	93	91	91	86	86
88	82	88	90	88	89
125	119	113	108	111	110
137	136	132	128	126	120
105	108	114	114	118	113
96	103	112	108	111	107

66	80	77	80	87	77
81	93	96	99	86	85
83	83	91	94	92	88
135	128	126	112	107	106
141	129	129	117	115	101
95	99	109	108	112	109
84	93	107	101	105	102

Figure.3. Color Images

JPG, JPEG: JPEG stands for Joint Photographic Experts Group .Jpeg format is mainly used for color photographs. It is not good with sharp edges and it tends to blur the image a bit. This format became trendy with the innovation of the digital camera. Digital cameras mostly download photos to our computer as a Jpeg format.

GIF: GIF stands for Graphics Interchange Format. This format is best suited for text, drawing line screen shots, animations and cartoons. Gif format is limited to total number of 256 colors or it can be less. It is mostly used for loading the fast web pages. It also helps to makes great banner and logo for different webpage. Different type of animated pictures is saved in GIF format. For example, the flashing banner would be saved as a Gif file format.

Benefits: It is supported mostly by all web browsers, it is very small file size, Easy to load, Benefit for Transparencies, and animations and Image maps

Downfalls: We can use only basic colors, Complex pictures look horrible, No details of images are allowed.

PNG:

PNG stands for Portable Networks Graphic. This is one of the best image formats; still it was not always well-suited with all web browsers and image software. This is the best image format to use for the website. It is also used for logo's and screen shots.

TIFF:

TIFF stands for Tagged Image File Format. This format has not been restructured since 1992 and is now owned by Adobe. It can store an image and data (tag) in the one file. This file is commonly used for scanning the data, faxing, word processing etc. It is no common file format that can be use with our digital photos.

Benefits: The image is perfect, Never loss any image.

Downfalls: Due to massive file size there is difficulty in transferring of the file, not able to view on the internet, only some specialized program can view it.

1.2 COLOR MODELS

For science communication, the two main colour spaces are RGB and CMYK.

1.2.1 RGB

The RGB colour model relates very closely to the way we perceive colour with the r, g and b receptors in our retinas. RGB uses additive colour mixing and is the basic colour model used in television or any other medium that projects colour with light. It is the basic colour model used in computers and for web graphics, but it cannot be used for print production. The secondary colours of RGB – cyan, magenta, and yellow – are formed by mixing two of the primary colours (red, green or blue) and excluding the third colour. Red and green combine to make yellow, green and blue to make cyan, and blue and red form magenta. The combination of red, green, and blue in full intensity makes white. In Photoshop using the "screen" mode for the different layers in an image will make the intensities mix together according to the additive colour mixing model. This is analogous to stacking slide images on top of each other and shining light through them.

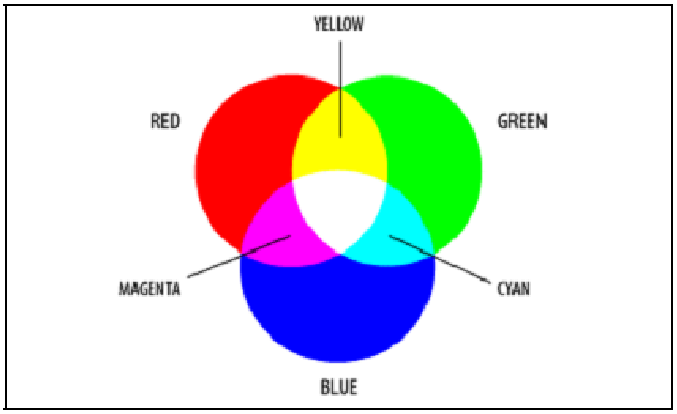


Figure.4. RGB Color Model

1.2.3 Gamut

The range, or gamut, of human color perception is quite large. The two color spaces discussed here span only a fraction of the colors we can see. Furthermore the two spaces do not have the same gamut, meaning that converting from one color space to the other may cause problems for colors in the outer regions of the gamuts.

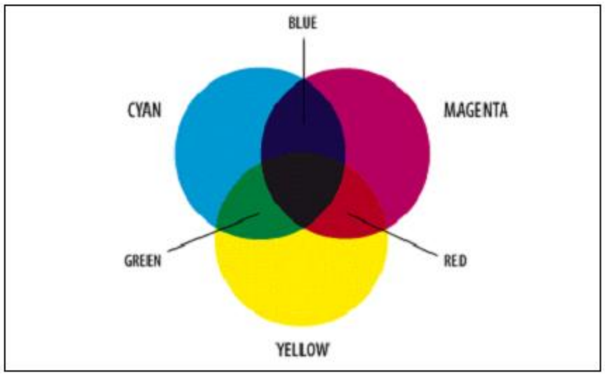


Figure.5. CMYK Color Model

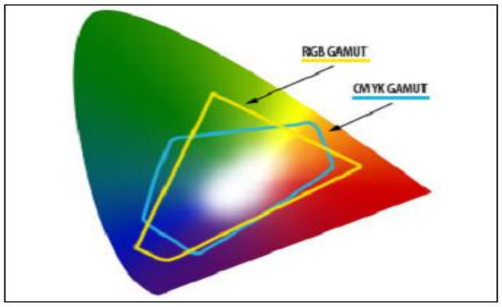


Figure.6. Different gamuts of the RGB and CMYK colour

1.3 CHARACTERISTICS OF IMAGE OPERATIONS

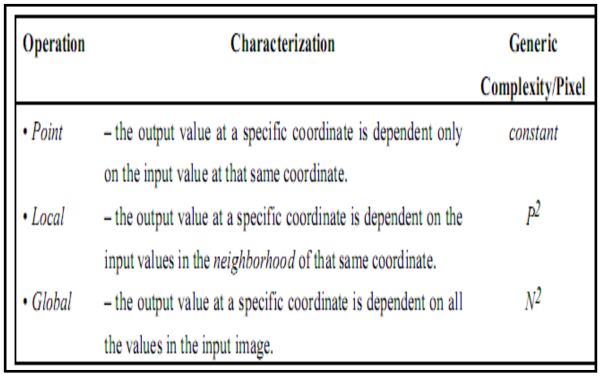
There is a variety of ways to classify and characterize image operations. The reason for doing so is to understand what type of results we might expect to achieve with a given type of operation or what might be the computational burden associated with a given operation.

1.3.1 Types of operations

The types of operations that can be applied to digital images to transform an input image a[m,n] into an output image b[m,n] (or another representation) can be classified into three categories as shown in Fig 1.8.

1.3.2 Types of neighborhoods

Neighborhood operations play a key role in modern digital image processing. It is therefore important to understand how images can be sampled and how that relates to the various neighborhoods that can be used to process an image.



Operation	Characterization	Generic Complexity/Pixel
• *Point*	– the output value at a specific coordinate is dependent only on the input value at that same coordinate.	*constant*
• *Local*	– the output value at a specific coordinate is dependent on the input values in the *neighborhood* of that same coordinate.	P^2
• *Global*	– the output value at a specific coordinate is dependent on all the values in the input image.	N^2

Figure.7. Image Operators

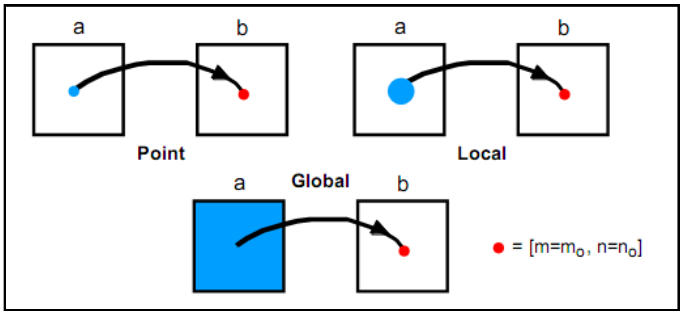


Figure.8. Illustration of various types of image operations

- Rectangular sampling – In most cases, images are sampled by laying a rectangular grid over an image as illustrated in Figure 1.10. This results in the type of sampling shown in Figure 1.10ab.

- Hexagonal sampling – An alternative sampling scheme is shown in Figure 1.11 and is termed hexagonal sampling.

Both sampling schemes have been studied extensively and both represent a possible periodic tiling of the continuous image space. Image will restrict our attention, however, to only rectangular sampling as it remains, due to hardware and software considerations, the method of choice. Local operations produce an output pixel value b[m=mo,n=no] based upon the pixel values in the neighborhood of a[m=mo,n=no]. Some of the most common neighborhoods are the 4-connected neighborhood and the 8-connected neighborhood in the case of rectangular sampling and the 6-connected neighborhood in the case of hexagonal sampling illustrated in Figure 3.

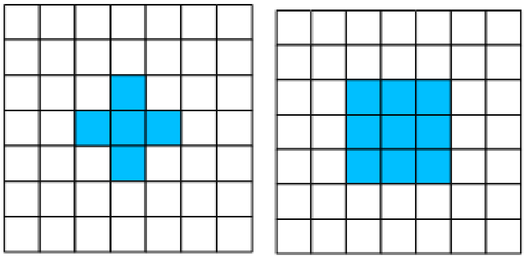


Figure.9. Rectangular sampling a and b

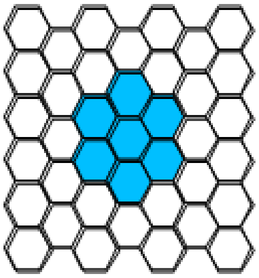


Figure.10. Hexagonal sampling

2.1 A COMPARATIVE STUDY ON MULTI-PERSON TRACKING USING OVERLAPPING CAMERAS

M. C. Liem and D. Et al [1] describes a comparative study for tracking multiple persons using cameras with overlapping views. The evaluated methods consist of two batch mode trackers (and one recursive tracker, which integrate appearance cues and temporal information differently. This paper [1] also added our own improved version of the recursive tracker. Furthermore, this paper investigate the effect of the type of background estimation (static vs. adaptive) on tracking performance. Experiments are performed on two novel and challenging multi-person surveillance data sets (indoor, outdoor), made public to facilitate benchmarking. This paper [1] shows that our adaptation of the recursive method outperforms the other stand-alone trackers. Tracking multiple persons in dynamic, uncontrolled environments using cameras with overlapping views has important applications in areas such as surveillance, sports and behavioral sciences. This paper [2] is interested in scenes covered by as few as surrounding cameras with diagonal viewing directions, maximizing overlap area. This set-up make s establishing individual feature correspondences across camera views difficult, while inter-person occlusion can be considerable. Various methods have been proposed recently for such a multi-person tracking setting using overlapping cameras, but few quantitative comparisons have been made. In order to improve visibility regarding performance characteristics, this paper present an experimental comparison among representative state-of-the-art methods. This paper [3] is selected one recursive method and two batch methods for this comparison. Furthermore, this made some performance improving adaptations to. The trackers were combined with the static background estimation method from and the adaptive background estimation method.

2.2EFFICIENTLY SCALING UP CROWD SOURCED VIDEO ANNOTATION

Carl Vondrick et al [2] presents an extensive three year study on economically annotating video with crowd sourced marketplaces. Our public framework has annotated thousands of real world videos, including massive data sets unprecedented for their size, complexity, and cost. To accomplish this, we designed a state-of-the-art video annotation user interface and demonstrate that, despite common intuition, many contemporary interfaces are sub-optimal. This paper present several user studies that evaluate different aspects of our system and demonstrate that minimizing the cognitive load of the user is crucial when designing an annotation platform. This paper then deploys this interface on Amazon Mechanical Turk and discover expert and talented workers who are capable of annotating difficult videos with dense and closely cropped labels. We argue that video annotation requires specialized skill; most workers are poor annotators, mandating robust quality control protocols. This paper [5] presents the traditional crowd sourced micro-tasks which are not suitable for video annotation and instead demonstrate that deploying time-consuming macro-tasks on MTurk is effective. Finally, it also show that by extracting pixel-based features from manually labeled key frames, this paper is able to leverage more sophisticated interpolation strategies to maximize performance given a fixed budget. This paper [6] show the results of three years of experiments and experience in annotating massive videos unprecedented for their size and complexity, with some data sets consisting of millions of frames, costing tens of thousands of dollars, and requiring up to a year of continuous work to annotate. This extensive study has resulted in our release of VTIC (Video Annotation Tool from Irvine, California), an open platform for monetized, high quality, crowd source video labeling. The contributions made in this paper are activated by our desire to uncover best-practices for monetized crowd sourced video labeling. In the remainder of this paper, this paper [7] describes our video annotation tool in detail: This paper briefly review related work in designing image and video annotation tools. This paper [6] present insights into the design of a user-interface in which workers track objects through a continuous video shot to support our claims, this paper present user studies that demonstrate contemporary annotation software is suboptimal. This paper [7] have found that video annotation is considerably more complex than mage annotation, likely due to the fact that temporal data is difficult to visualize and edit. This paper describes how to best use crowd sourcing to annotate videos. In order to collect high quality annotations, this paper finds it crucial to validate good workers and turn away the majority of MTurk workers. In this sense, we do not use MTurk directly as a crowd sourced platform, but rather as a market to identify reliable workers. This paper analyze trade-offs particular to balancing computer and human effort in video annotation by extending work that minimized labeling cost only along the dimension of human effort. Although the "Turk philosophy" is to completely replace difficult computer tasks (such as video labeling) with human effort, this is clearly not efficient given the redundancy of video. In contrast to Label Me ideo, this paper [7] show that one can interpolate nonlinear least-cost paths with efficient dynamic programming algorithms based on image data and user annotated endpoints. This paper analyzes the total cost of labeling for various combinations of human workers and cloud computing CPU cycles. This paper further demonstrate that our cost analysis can be used as an error metric for evaluating vision algorithms; rather than evaluating a tracker with disconnected measures such as time-to-failure. The red boxed player becomes totally occluded while many players quickly change pose from standing to a prone position. The referees commonly enter and leave the scene. The camera is not stationary. The ball exists in the pile of people, but even a state-of-the-art vision algorithm is unable to determine its position. Evaluate trackers using the dollar amount in savings afforded when used in a monetized, interactive crowd sourced platform. Our hope is that our discoveries will spur innovation in the creation of affordable, massive data sets of labeled video. To encourage this, our final contribution is the release of a simple, reusable, and open-source platform for research video labeling. This paper [8] has introduced a large scale video annotation platform capable of economically obtaining high quality labels for complex videos. This paper first built an efficient user interface for video annotation by informing our de sign choices through extensive user studies. Our

2.3 A MULTI-VIEW ANNOTATION TOOL FOR PEOPLE DETECTION EVALUATION

Ákos Utasi et al [3] introduces a novel multi-view annotation tool for generating 3D ground truth data of the real location of people in the scene. The proposed tool allows the user to accurately select the ground occupancy of people by aligning an oriented rectangle on the ground plane. In addition, the height of the people can also be adjusted. In order to achieve precise ground truth data the user is aided by the video frames of multiple synchronized and calibrated cameras. Finally, the 3D annotation data can b e easily converted to 2D image positions using the available calibration matrices. One key advantage of the proposed technique is that different methods can b e compared against each other, whether they estimate the real world ground position of people or the 2D position on the camera images. Therefore, this paper defined two different error metrics, which quantitatively evaluate the estimated positions. This paper [9] used the proposed tool to annotate two publicly available datasets, and evaluated the metrics on two state of the art algorithms. In many surveillance systems key functionalities involve pedestrian detection and localization in the scene. The location information is used in higher level modules, such as tracking, people counting, restricted zone monitoring, or behavior analysis. In recent years multi-view surveillance has undergone a great advance, and novel methods have been proposed to improve the efficiency of people detection and localization. However, most of existing multi-view image sequences is annotated using the conventional method of generating 2D bounding boxes around the pedestrians in the images. This work presents a novel approach for manual ground truth generation for multi-view image sequences and goes beyond the traditional bounding box annotation technique. The proposed tool assumes that the multiple sequences are synchronized and the cameras are calibrated. Moreover, an area of interest (AOI) is also defined by the user on the ground plane. In the proposed annotation the real 3D ground position, the occupancy area on the ground plane (represented by oriented rectangles), and the height is stored for each person. The rest of the paper is organized as follows. This paper [10] also gives a brief overview of existing annotation datasets for the people surveillance. In the proposed ulti-view annotation tool is presented. Moreover, in this section we briefly present two public datasets we manually annotated with our tool. In Sec. 4 we present two different error metrics which can be used to evaluate pedestrian localization algorithms using our annotation data.

2.4 IS MY NEW TRACKER REALLY BETTER THAN YOURS

Luka Cehovin et al [4] describe a visual tracking evaluation is sporting an abundance of performance measures, which are used by various authors, and largely suffers from lack of consensus about which measures should be preferred. This is hampering the cross-paper tracker comparison and faster advancement of the field. In this paper [13] provide an overview of the popular measures and performance visualizations and their critical theoretical and experimental analysis. We show that several measures are equivalent from the point of information they provide for tracker comparison and, crucially, that some are more brittle than the others. Based on our analysis we narrow down the set of potential measures to only two complementary ones that can be intuitively interpreted and visualized, thus pushing towards homogenization of the tracker evaluation methodology.
Visual tracking is one of the rapidly evolving fields of computer vision. Every year, literally dozens of new tracking algorithms are presented and evaluated in journals and at conferences. When considering the evaluation of these new trackers and comparison to the state-of-the-art, several questions arise. Is there a standard set of sequences that we can use for the evaluation? Is there a standardized evaluation protocol? What kind of performance measures should we use? Unfortunately, there are currently no definite answers to these questions. Unlike some other fields of computer vision, like object detection and classification, optical-flow computation and automatic segmentation, where widely adopted evaluation protocols are used, visual tracking is still largely lacking these features. The absence of homogenization of the evaluation protocols makes it difficult to rigorously compare trackers across publications and stands in the way of faster development of the field. The authors of new trackers typically compare their work against a limited set of related algorithms due to the difficulty of adapting these for their own use in the experiments [14]. The issue here is the choice of tracker's performance evaluation measures, which seems to be almost arbitrary in the tracking literature. Worse yet, an abundance of these measures are currently in use. Because of this, experiments in many cases offer a limited insight into tracker's performance, and prohibit comparison across different papers. In this paper [15] focus on the problem of performance evaluation in monocular single-target visual tracking and address several challenges therein. The goal of this paper is not to propose new performance measures. Instead we focus on narrowing the wide variety of existing measures for single-target tracking performance valuation to only a few complementary ones. This is a crucial step towards the homogenization of the field f is. To claim a three-fold contribution: to provide a detailed survey and

experimental analysis of performance measures used in single-target tracking evaluation. It shows by experimental analysis that there exist clusters of performance measures that essentially indicate the same aspect of tracker's performance. By considering the theoretical aspects of existing measures as well as the experimental analysis we identify the two most suitable complementary measures that characterize tracker's performance within the accuracy vs. robustness context and propose an intuitive way to visualize the selected pair of measures.

2.5 PERFORMANCE EVALUATION OF MULTI-CAMERA VISUAL TRACKING

Lucio marcenaro et al [5] describe an in single-camera multi-target visual tracking can be partially removed by ncreasing the amount of information gathered on the scene, i.e. by adding cameras. By adopting such a multi-camera approach, multiple sensors cooperate for overall scene understanding. However, new issues arise such as data association and data fusion. This work [17] addresses the issue of evaluating the performance of a multi camera tracking algorithm based on Rao-Black wellized Monte Carlo data association (RBMCDA) on real data. For this purpose, a new metric based on three performance indexes is developed.

Video-based tracking techniques are widely used in different fields related to ambient intelligence and scene understanding. Video surveillance is only one of the possible applications ranging from security to man-machine interaction; elderly people monitoring for sanitary assistance or rehabilitation is only one of the interesting fields where video-sensors and automatic processing algorithms can be used efficiently. Moreover, technological improvements in imaging hardware and related price cuts enabled not even more pervasive usage of such sensors: IP based video cameras allowed easy hardware installation and remote monitoring and configuration while mega-pixel sensors and efficient video compression algorithms greatly improved quality and spatio-temporal resolution of acquired images. In this scenario [18], visual tracking is one of the principal techniques that are used for extracting high-level descriptors of the scene. The main purpose of tracking algorithms is to estimate moving objects' trajectories by correctly assigning them a specific label that must be propagated over-time as the object moves on the scene. Each tracking technique starts from low-level data acquired directly from images, thus can be modeled as a state estimate problem starting from noisy observations. In this sense [19] tracking can be seen as a filtering process for canceling observations noise and extracting the actual movement of each object in the scene. Several high-level applications can be developed based of trajectories and detected movements, such as people counting, traffic monitoring, abandoned objects detection, automotive safety, etc. The complexity of visual-tracking is very high and main problems that must be solved are related with unpredictable object movements, targets or scene appearance variations, no rigid parts, occlusions (intra-, inter- or environmental) and sensor undesired movements. A multi-camera approach can be adopted in order to deal with these issues, although data fusion and association problems must be considered when multiple sensors have to cooperate for overall scene understanding. Visual tracking problems can be mathematically modeled through probabilistic reasoning. Target positions can be stored in a state vector containing its descriptive features: target state must be estimated starting from noisy observations obtained from acquired images. Stochastic state and observation models are needed for this kind of approach. Bayesian theory can be used to estimate the state's a posteriori probability density function (pdf): after the pdf is estimated, it can be used for evaluating the new state and for measuring the state estimate precision. Performances have been evaluated by tuning some of the parameters introduced in section II-B, namely Pb, which models the birth probability density, and the two parameters α and β, which shape the Gamma function that models the lifetime probability of each target. Two 500-frame scenes with a different intrinsic complexity have been studied. The will be referred to as Scene 1 and 2 (simpler and more complex scene respectively). The three parameters have been varied in turn, while keeping the two other fixed. Summarizes. The ranges of α and β and *Pb* have been explored schematically to localize local maxima for the tree performance indexes. Low values of α result in short lifetimes, and in fact in trajectories fragmentation, as one can infer from the positive values of *IF* .On the other hand, the values of *IT* and *IV* do not show monotone city; instead there seems to be a minima at $\alpha = 4$. By varying β , the fragmentation *IF* remains constant at the optimal null value, while the other two indexes' behavior is quite fuzzy. $\beta = 1$ seems to give the best fit. Eventually (Table I(c)), it can be noticed that better performances are obtained with the higher considered value of *Pb*, i.e. $Pb = 0.15$. *IV* and *IT* improve as *Pb* rises: this is basically due to the fact that there is less delay in hooking the target as the birth probability increases. Even higher values of Pb result in positive values of *IF* (as for Pb = 0.1), as the algorithm generates too many targets. Summarizes some results for *Scene 1*. Here scene is complicated by more targets, resulting in more occlusions and misdetections. The optimal value for α has been found for $\alpha = 8$. It can be noticed that the value of the three indexes show here a kind of local minimum: in particular, the algorithm estimates more trajectories than the real ones for $\alpha < 7$ ($IF > 0$) and less for $\alpha > 8$ ($IF < 0$). Table II(b) shows how $\beta = 5$ represents some kind of asymptote turning point for the values of all the indexes. in particular *IV* turns positive, meaning that too delayed target deaths result in too many people detected on the scene. From the analysis of Table II(c), it looks like that than algorithm cannot avoid fragmenting trajectories in a more complex scene. A set-up for RBMCDA testing has been presented [20]. A novel metric based on three indexes has been exploited in order to evaluate the algorithm performance, both qualitatively and quantitatively. Each index accounts for a typical problem in visual tracking. Performance has been evaluated as α, β and Pb, which model target birth and death probabilities, vary. Future developments of this work [19] surely include comparative performance analysis with other trackers. Including simple particle filter (non-marginalized filter) evaluation could give evidence to the Rao-Blackwell theorem

3.1 ECGM-BASED

- While continuous measures of the strength of relationship hold complete information, but it is highly sensitive to noises.
- Same person name for given face is tracked even if movies vary.
- Noise removal process is not discussed.
- The sequential statistics for the speakers is not carried out.

3.1.1 DATA INPUT

- Movie files selection using open file dialog control. AVI File is selected as input and saved in table.
- Movie file is selected from table, and split into individual frames using AVI extractor “avifil32.dll” methods and saved as bitmaps. The bitmaps folder in the project is used to save all the frames. The record is saved into ‘Bit maps’ folder with movie id and frame id.
- Movie file is selected from table, frame id is selected from the retrieved bitmap frame id is selected and title sentence is added.
- Person names found in title is added into ‘Face Names’ table with movie id, frame id and name.

3.1.2 FACE ANNOTATION

- After the movie id selection, from the bitmap frames face area is found out and the details are saved in ‘Faces’ table with movie id, frame id and face data.
- From the frames, each one is selected.
- Converted into gray scale image
- Morphological filter is applied with erosion property (3X3) matrix is given as input for erosion process.
- Then Contour (Border) is found out. Then based on the given width/height ratio, places where images can be found out.
- Select face regions’ Image data are saved into database with X and Y location along with width and height of the area.

3.1.3 FACE CLUSTERING WITH ANNOTATION

- After the movie id selection, faces are clustered such that K Means clustering is applied with ‘N’ clusters is given as input.
- Based on the color difference in the bitmap pixels, the face similarity is calculated.

3.1.4 FORMING FACE RECOGNITION

- Based on the faces (multiple person) appeared in the bitmap frames, relationship between faces is formed.
- For example, one frame contains Face A, B and C other contains A and C. So A is more related with C and less related B.
- The edge weight is fixed based on the relationship/ common occurrence between faces.

3.1.5 FORMING PERSON ANNOTTAION GRAPH

- Based on the names (multiple person named) appeared in the bitmap frames title, relationship between names is formed.
- For example, one frame contains name A, B and C other contains A and C. So A is more related with C and less related B.
- The edge weight is fixed based on the relationship/ common occurrence between names.

3.1.6 PERSON RECOGNTION

- Matching is done based on common occurrences of faces and names.
- For example If frame 1 contains Face A, B and C with name X, Y and Z.
- Then frame 2 contains Face A and C with name X and Z, then it is sure that A and B have the names X and Z.
- After intersecting all the frames with Face/Person occurrence, try to match the person with faces.

4. CONCLUSION

A methodology for generating reference tracking data in long multi-camera videos, based on the consensus of a detector and several trackers. For multi-camera annotation, our methodology is the first to estimate the reliability of annotations, and to offer the possibility of balance accuracy and human-effort in the final annotation result. A novel probabilistic framework is to learn the error models of trackers, and how to apply them to estimate target position. Previous methods did not model the tracking error statistics of multiple trackers to increase the reliability of the estimated position. The valuation of the accuracy and reliability of the proposed methodology 6 hour dataset, by a comprehensive visual inspection. Scalability of a semi-automatic methodology for annotation in multi-camera data sets of such length is addressed for the first time.

5. REFERENCES

[1].A. Milan, K. Schindler, and S. Roth, “Challenges of ground truth evalua-tion of multi-target tracking,” in Proc. CVPRW, Jun. 2013, pp. 735–742.

[2].H. Wu, A. C. Sankaranarayanan, and R. Chellappa, “Online empirical evaluation of tracking algorithms,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 8, pp. 1443–1458, Aug. 2010.

[3].M. Kristan et al., “The visual object tracking VOT2013 challenge results,” in Proc. ICCVW, Dec. 2013, pp. 98–111.

[4].M. Kristan et al., “The visual object tracking VOT2014 challenge results,” in Proc. ECCV Workshops, 2014, pp. 191–217.

[5].J. Yuen, B. Russell, C. Liu, and A. Torralba, “LabelMe video: Building a video database with human annotations,” in Proc. ICCV, Sep./Oct. 2009, pp. 1451–1458.

[6].B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, “LabelMe: A database and Web-based tool for image annotation,” Int. J. Comput. Vis., vol. 77, nos. 1–3, pp. 157–173, 2008.

[7].A. Dorado, J. Calic, and E. Izquierdo, “A rule-based video annotation system,” IEEE Trans. Circuits Syst. Video Technol., vol. 14, no. 5, pp. 622–633, May 2004.

[8].M. Bertozzi, A. Broggi, R. Grisleri, A. Tibaldi, and M. D. Rose, “A tool for vision based pedestrian detection performance evaluation,” in Proc. IEEE Intell. Vehicles Symp., Jun. 2004, pp. 784–789.

[9].C. Spampinato, B. J. Boom, and J. He, “MTAP special issue on meth-ods and tools for ground truth collection in multimedia

applications," Multimedia Tools Appl., vol. 70, no. 1, pp. 409–412, May 2014.

[10].S. Vijayanarasimhan and K. Grauman, "Active frame selection for label propagation in videos," in Proc. ECCV, 2012, pp. 496–509.

[11].V. Karasev, A. Ravichandran, and S. Soatto, "Active frame, location, and detector selection for automated and manual video annotation," in Proc. CVPR, 2014, pp. 2123–2130.

[12].I. Kavasidis, S. Palazzo, R. Di Salvo, D. Giordano, and C. Spampinato, "A semi-automatic tool for detection and tracking ground truth genera-tion in videos," in Proc. VIGTA, 2012, pp. 1–5.

[13].I. Kavasidis, S. Palazzo, R. D. Salvo, D. Giordano, and C. Spampinato, "An innovative Web-based collaborative platform for video annotation," Multimedia Tools Appl., vol. 70, no. 1, pp. 413–432, May 2013.

[14].C. Vondrick and D. Ramanan, "Video annotation and tracking with active learning," in Proc. NIPS, 2011, pp. 28–36.

[15].F. Comaschi, S. Stuijk, T. Basten, and H. Corporaal, "A tool for fast ground truth generation for object detection and tracking from video," in Proc. ICIP, Oct. 2014, pp. 368–372. J. Kwon and K. M. Lee, "Tracking by sampling trackers," in Proc. ICCV, Nov. 2011, pp. 1195–1202.

[16].Q. Li, X. Wang, W. Wang, Y. Jiang, Z.-H. Zhou, and Z. Tu, "Disagreement-based multi-system tracking," in Proc. ACCV Workshop Detection Tracking Challenging Environ., 2012, pp. 320–334.

[17].C. Bailer, A. Pagani, and D. Stricker, "A superior tracking approach: Building a strong tracker through fusion," in Proc. ECCV, 2014, pp. 170–185.

[18]. Y. Gao, R. Ji, L. Zhang, and A. Hauptmann, "Symbiotic tracker ensemble toward a unified tracking framework," IEEE Trans. Circuits Syst. Video Technol., vol. 24, no. 7, pp. 1122–1131, Jul. 2014.

[19].J. Whitehill, T.-F. Wu, J. Bergsma, J. R. Movellan, and P. L. Ruvolo, "Whose vote should count more: Optimal integration of
[20].N. Wang and D.-Y. Yeung, "Ensemble-based tracking: Aggregating crowd sourced structured time series data,"