How to be an Effective and Successful Graduate Student

Jenq-Neng Hwang, Professor

Associate Chair for Global Affairs

Department of Electrical Engineering

University of Washington, Seattle WA

hwang@uw.edu

www.ee.washington.edu/faculty/hwang/

Important Features (M^2I^4) Required for Graduate Study

- Motivation: really enjoy digging deeper into the scientific truth
- Maturity: never afraid of being failed or left alone
- Innovation: always think of what is (or can be) new and different?
- Intelligence: filter useful information to become usable knowledge
- Independence: a step-by-step problem formulation and module solving
- Integrity: always be honest in reporting the results

Getting Ready for Research

- Research topics selection and switching
  - Extend from senior students' topics
  - Existing research projects in the Lab
  - Something your advisor is willing to learn closely with you
- **Depth Knowledge:**
  - A good series of class taking, or self study related tutorial background
  - Most updated Conference/Journal papers (IEEE Xplore)
  - Joint project discussions and group collaborations, or new class offering
- **Breadth Knowledge:**
  - Attend technical presentation and active questioning (key messages?)
  - Magazine and Hi-Tech News
- How much paper citations can tell you about your chosen topic?

Google Scholar

User profiles for **jenq neng hwang**

Jenq-Neng Hwang
University of Washington
Verified email at u.washington.edu
Cited by 8108

[BOOK] Handbook of neural network signal processing
YH Hu, JN **Hwang** - 2001 - books.google.com
The use of neural networks is permeating every area of signal processing. They can provide powerful means for solving many problems, especially in nonlinear, real-time, adaptive, and blind signal processing. The Handbook of Neural Network Signal Processing brings
Cited by 422 Related articles All 5 versions Cite Save More

Fast and automatic video object segmentation and tracking for content-based applications
C Kim, JN **Hwang** - IEEE transactions on circuits and systems …, 2002 - ieeexplore.ieee.org
Abstract: The new video-coding standard MPEG-4 enables content-based functionality, as well as high coding efficiency, by taking into account shape information of moving objects. A novel algorithm for segmentation of moving objects in video sequences and extraction of
Cited by 467 Related articles All 8 versions Cite Save

Nonparametric multivariate density estimation: a comparative study
JN **Hwang**, SR Lay, A Lippman - IEEE Transactions on Signal …, 1994 - ieeexplore.ieee.org
Abstract: The paper algorithmically and empirically studies two major types of nonparametric multivariate density estimation techniques, where no assumption is made about the data being drawn from any of known parametric families of distribution. The first type is the
Cited by 260 Related articles All 12 versions Cite Save

Lipreading from color video
GI Chiou, JN **Hwang** - IEEE Transactions on Image Processing, 1997 - ieeexplore.ieee.org
Abstract: We have designed and implemented a lipreading system that recognizes isolated words using only color video of human lips (without acoustic data). The system performs video recognition using" snakes" to extract visual features of geometric space, Karhunen-
Cited by 162 Related articles All 7 versions Cite Save

Fast and automatic video object segmentation and tracking for content-based applications

Search within citing articles

Video shot detection and condensed representation. a review
C Cotsaces, N Nikolaidis, I Pitas - IEEE signal processing …, 2006 - ieeexplore.ieee.org
Abstract: There is an urgent need to develop techniques that organize video data into more compact forms or extract semantically meaningful information. Such operations can serve as a first step for a number of different data access tasks such as browsing, retrieval, genre
Cited by 294 Related articles All 7 versions Cite Save

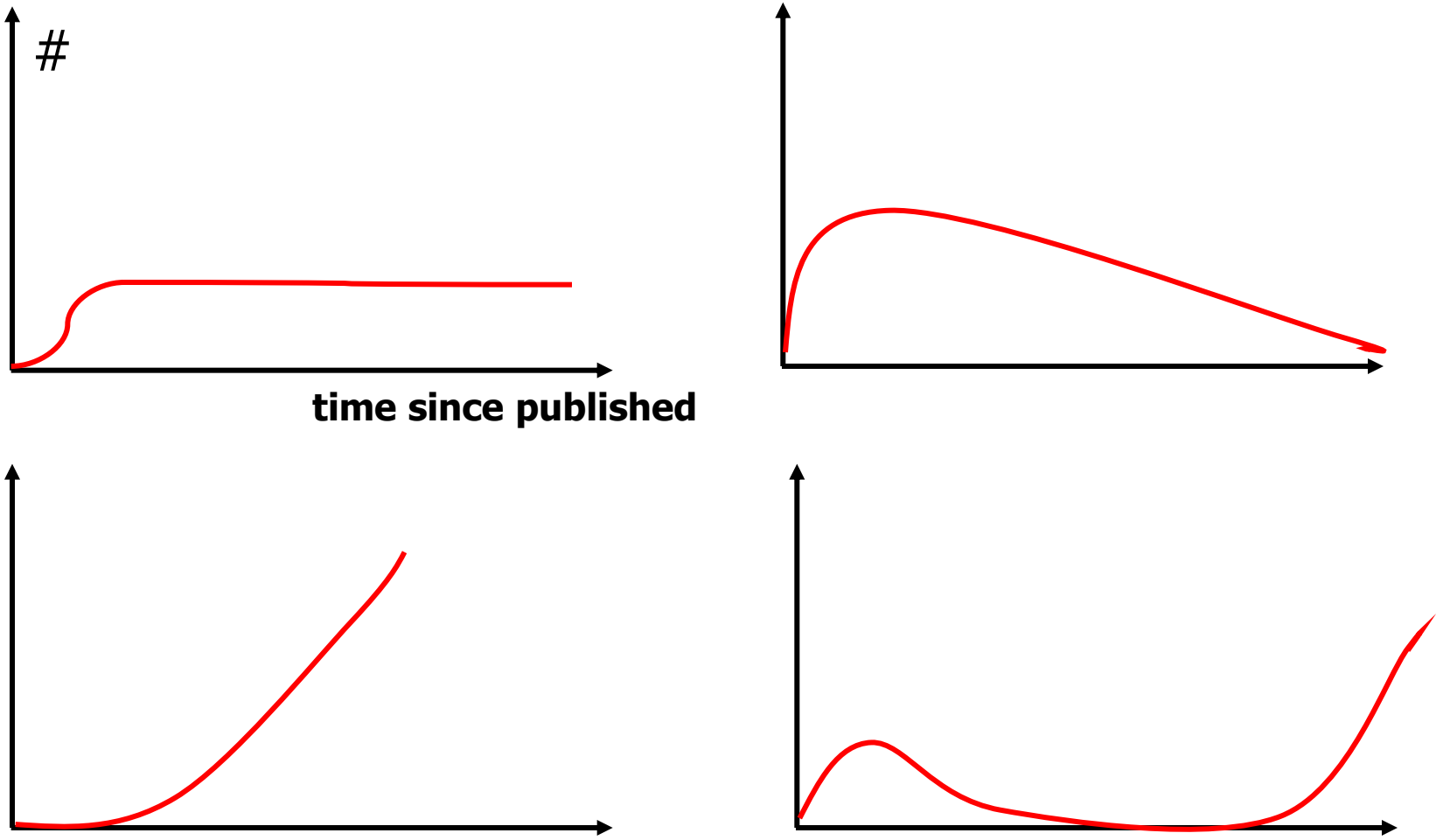
Video object tracking using adaptive Kalman filter
SK Weng, CM Kuo, SK Tu - Journal of Visual Communication and Image …, 2006 - Elsevier
In this paper, a new video moving object tracking method is proposed. In initialization, a moving object selected by the user is segmented and the dominant color is extracted from the segmented target. In tracking step, a motion model is constructed to set the system
Cited by 213 Related articles All 7 versions Cite Save

3D motion retrieval with motion index tree
F Liu, Y Zhuang, F Wu, Y Pan - Computer Vision and Image Understanding, 2003 - Elsevier
With the development of Motion capture techniques, more and more 3D motion libraries become available. In this paper, we present a novel content-based 3D motion retrieval algorithm. We partition the motion library and construct a motion index tree based on a
Cited by 190 Related articles All 13 versions Cite Save

Video browsing by direct manipulation
P Dragicevic, G Ramos, J Bibliowitcz… - Proceedings of the …, 2008 - dl.acm.org
Abstract We present a method for browsing videos by directly dragging their content. This method brings the benefits of direct manipulation to an activity typically mediated by widgets. We support this new type of interactivity by: 1) automatically extracting motion data from
Cited by 130 Related articles All 23 versions Cite Save

Object-based video abstraction for video surveillance systems

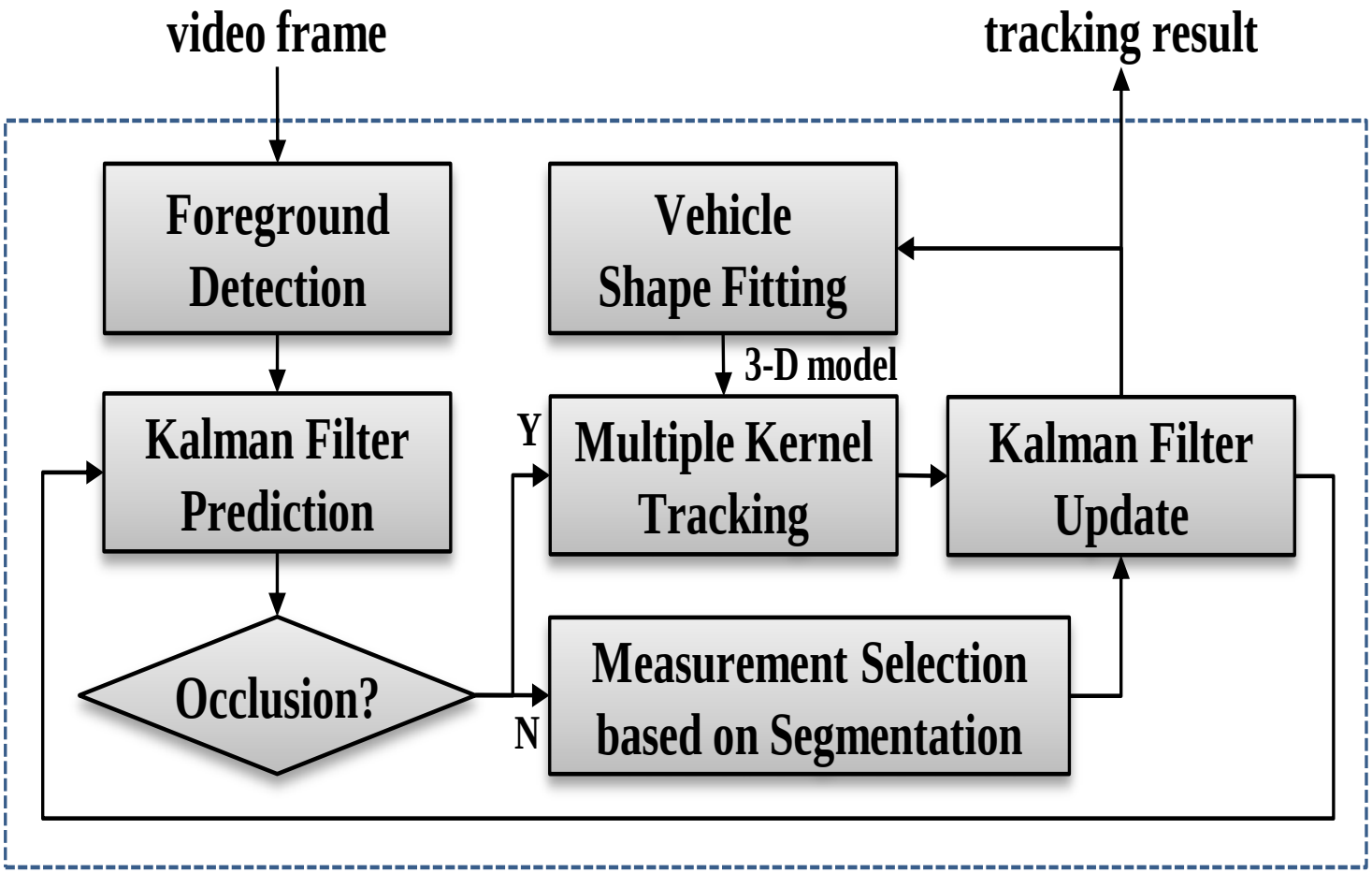
Paper Citations (e.g., Google Scholar)



Jumping into Research

- Extensive literature survey and problem formulation
- Summarize others' paper in their block diagrams (flow charts) – look for weak links or not convincing blocks
- Conclusion meet the original problem formulation?
- Move on to next block by interpreting and verifying the complete ideas
- Always use the data and compare the results with the most recent or best reported results
- Clear and detailed interpretation of simulation results
- Confidence and leadership building (organized speech in group or individual meetings), time management too

A Typical Block Diagram



Persistence in Research

- Never expect a smooth path, refine the research scopes all the time (backoff slightly to find brighter road ahead)
- Learn from any failure -- is a step closer to your final success (make you appreciate and hang on what you own and any small progress – JK Rowling 2008)
- Never hide from your advisor/teammate, even no/little progress (learn from mistakes or failures)
- Keep on thinking the block diagrams and see what can be further modified or improved
- Convince your advisor or your teammate your findings
- Research Training – completing a well planned flow chart (a manager/leader, not employee, training)

Research Publications and Technical Reports

- Practice English writing of thesis and papers: practice by mimicking
- Always seek publication opportunities: from conferences to periodical journals
- Learn the standard writing style and outline of manuscript (a good outline, 70% done!)
- Learn from the grammatical errors corrected by your advisor or technical editing persons
- Discuss clearly the flow charts and simulation results in your papers, never leave the weak links
- Never be afraid/frustrated of major/minor revision of paper submission – reviewers are never your enemies

Learn from Track Changes

training stage. Further, we employ the asymmetric bidirectional transition time distribution. ~~Due to the irreversible property of time distribution, it~~which is essential in improving the tracking performance of the camera networks~~link models~~.

The rest of this paper is organized as follows. In Section 2, we give an overview the overall tracking system. We provide the details of ~~demonstrate~~ the proposed algorithm in Section 3. The experimental results are conducted in Section 4, followed by ~~. Finally, we~~the conclusion~~s and summarize the paper~~ in Section 5.

2. SYSTEM OVERVIEW

~~Our~~The original unsupervised multi-camera tracking system proposed in [3] ~~is~~ consists of two parts as the following (see Fig. 1):

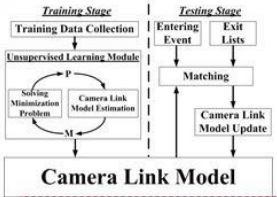


Fig. 1: A multiple camera tracking system [3].

2.1. Training Stage

First, ~~humana~~ person who leaves or enters the field of view (FOV) of ~~eacha~~ pair of exit/entry zones in two directly-connected cameras are collected [3~~4~~]. The deterministic annealing and the barrier method~~s~~ are applied to estimate the "symmetrical" camera link model between two directly-connected ~~cameras~~ ~~zones~~subsequently. The camera link model contains several features: transition time distribution, brightness transfer functioning (BTF) [5], region mapping matrix, region matching weight, and feature fusion weight [3]. These results are used on the following testing stage, assuming the same camera link model can be applied on either direction of human movement, to re-identify the human who cross the cameras.

2.2. Testing Stage

All the multiple cameras perform the single camera human tracking by adaptive Kalman filtering and multiple kernels tracking with projected gradients [4]. Each camera C_i, $i = 1 \sim N_C$, preserves an exit list $\mathbf{L}_{i,k}$ for each exit/entry zone k,

$$\mathbf{L}_{i,k} = \left\{ O^1_{i,k}, O^2_{i,k}, \ldots, O^{N_{i,k}}_{i,k} \right\}, \quad (1)$$

~~The list has~~which contains the observations of the person~~ople~~ who have left the FOV from zone k within training time T_{max} seconds from now. When a person enters FOV of the other connected camera, the tracking algorithm across cameras finds the best correspondence among the exit lists by using the matching score. The observation which gets the low score according to Eq. (2) is decided to the best correspondence where α_{ij} is the weight of feature j~~i~~ derived from the training stage.

$$score = -\sum_{j=1}^{N_{feature}} \alpha_j \times feature\_dist_j \quad (2)$$

When the score is higher than the predefined threshold, we judge that the best correspondence is the same exit person~~human~~ from the exit camera C_i. Otherwise, we will regard it as a new person ~~human~~. The re-identification results can be further used to update the camera link models.

3. PROPOSED ALGORITHMS

We propose ~~athe~~ new estimation method to produce more reliable camera link models for camera networks. Through combining several~~plural~~ links in the training stage, each link ~~becomes to have each~~can produce better camera link model ~~which shows the relationship~~ between each directly connected camera pair. In this paper, we ~~used~~ jointly train ~~the pairwise camera link models based on a group of the~~ three cameras, ~~each of~~ which ha~~s~~ve two different connected links ~~to explain for the~~our proposed algorithms. The deterministic annealing is again employed [6]~~[7]~~ to build the optimum binary permutation matrix $\mathbf{P}$ and the corresponding camera link model between each pair of connected cameras.

3.1. Camera Link Model Estimation

In the training stage, we estimate the camera link model based on~~including several components with~~ sets of observations acquired from directly-connected cameras. As shown in ~~In~~ Fig. 2~~1~~, where camera C_1 is directly connected to camera C_2~~s~~. ~~Suppose we have given two the exit~~ sets, $\mathbf{X}$ and the entry set $\mathbf{Y}$, identified from ~~exit, C_1, and entry~~ and C_2, respectively.

$$\mathbf{X} = \left[x_1, x_2, \ldots, x_{N_1} \right], \quad \mathbf{Y} = \left[y_1, y_2, \ldots, y_{N_2} \right], \quad (3)$$

where x_i and y_j are exit and entry observations, and N_1 and N_2 are the corresponding number of the observations. Each element of $\mathbf{X}$ and $\mathbf{Y}$ ha~~s~~ve the exit or entry time stamp, holistic color, region color and texture features. We exploit these information to match the correspondences automatically between the exit/entry observation sets ~~of entry/exit~~. Our goal is to build the $(N_1+1) \times (N_2+1)$ permutation matrix $\mathbf{P}_{n\times}$. ~~The~~whose element P_{ij} ~~in matrix P~~ is set to 1 if x_i corresponds to y_j. Otherwise, it is set to 0. Finally the camera link model estimation process produces $(N_1+1) \times (N_2+1)$ correspondence matrix $\mathbf{P}$ and the extra~~and of~~ column and row represent the outlier entries. The problem can be written as a constrained minimization integer programming problem:

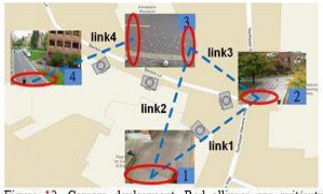


Figure ~~1~~2. Camera deployment. Red ellipses are exit/entry zones for four links and the number in blue rectangular means the camera number.

$$\hat{\mathbf{P}} = \arg\min_{\mathbf{P}} J(\mathbf{P}), \quad (4)$$

$$s.t.\ P_{ij} \in \{0,1\} \quad \forall i \le N_1+1, j \le N_2+1 \quad (5)$$

$$\sum_{i=1}^{N_1+1} P_{ij} = 1 \quad \forall j \le N_2, \quad \sum_{j=1}^{N_2+1} P_{ij} = 1 \quad \forall i \le N_1 \quad (6)$$

$$J(\mathbf{P}) = cost_{time} + cost_{holistic\,color} + cost_{region\,color} + cost_{region\,texture} + cost_{entropy} + cost_{outlier}, \quad (7)$$

where J is the objective function to be minimized and ~~cost~~$_{texture}$ is the cost function of each feature [3]. The objective function J ~~is going to be~~ is iteratively minimized with $\mathbf{P}$ approach~~ed~~ to binary matrix. Then converged matrix $\mathbf{P}$ is ~~closer to ground truth, we can~~then used to derive a ~~get~~ more reliable camera link model.

3.2. Combined Camera Network

Nowadays more and more~~many~~ surveillance cameras are ~~distributed~~getting deployed on the street~~s~~, resulting in a spider web like ~~the~~ connected camera ~~links, models being likeas a meshed spider web which has many meshes~~. ~~In case of~~ Most existing ~~previous~~ methods which use feature transfer function [1]~~[2]~~[3], ~~they~~ build link model by using 1-to-1 pair of cameras to represent the space-time relationship and color/texture transfer model~~s~~. However, if there are several forked road, many persons who leave one exit will not necessarily show up in one specific entry. They are going to be regarded as outlier in matrix $\mathbf{P}$. The error increases as the portion of outliers increases~~raises~~ [6]. In other words, the estimation of $\mathbf{P}$ may make more incorrect~~take more wrong~~ correspondences, resulting in, ~~and the product of estimation process,~~less accurate camera link model~~s, will be less accurate~~. Thus, we ~~propose~~are motivated to combine the several pairs of connected cameras ~~to~~during the estimation of the P_{ij}'s to jointly create ~~make~~more reliable camera link models.

As shown ~~1~~in Fig. 2~~1~~, which shows an actually deployed 4-camera network $\{C_1, C_2, C_3, C_4\}$ surrounding Electrical Engineering building of the University of Washington, there are in total four camera links ~~which areas~~ denoted by~~as~~ blue dash line~~s~~. Among ~~these 4 cameras~~~~of them~~, C_1~C_3 have ~~an~~ entry/exit~~,~~ zones which are~~is~~ directly connected to two different ~~way directly~~zones from two other cameras. So the person who exits from C_1 can enter to C_2 or C_3 by crossing the link1 or 2, vice versa for C_2 and C_3. According to our previous method [3], ~~they~~which disregard~~s~~ the person who enters ~~to~~ C_3 ~~when they~~during the building of the camera link model ~~of~~ for link1~~,~~. ~~Then~~ by treating this person ~~as~~is outlier. On the other hand, ~~B~~by combining link1 and link2, this person ~~changes to~~becomes an inlier, resulting in a lower ~~and the~~ outlier percentage ~~will decrease~~.

Suppose we have an additional set, $\mathbf{Z}$, from entry, C_3.

$$\mathbf{Z} = \left[z_1, z_2, \ldots, z_{N_3} \right], \quad (8)$$

where ~~z~~$_k$ is an entry observation~~s~~ and N_3 are the total number of observations within the training time T_{max}. By adding $\mathbf{Z}$ to $\mathbf{P}$, the problem can be rewritten as follows:

$$\hat{\mathbf{P}}' = \arg\min_{\mathbf{P}'} J(\mathbf{P}'), \quad (9)$$

$$-s.t.\ P'_{ij} \in \{0,1\}, \quad \forall i \le N_1+1, j \le N_2+N_3+1,$$

$$(10)$$

$$\sum_{i=1}^{N_1+1} P'_{ij} = 1 \quad \forall j \le N_2+N_3, \quad \sum_{j=1}^{N_2+N_3+1} P'_{ij} = 1 \quad \forall i \le N_1,$$

$$(11)$$

where $\mathbf{P}'$ is a revised concatenated matrix of $\mathbf{P}$. The wrong matches are occu~~rred~~ when the solver is trapped in the local minimum during solving the minimization problem with deterministic annealing [6]. However, $\mathbf{Z}$ contains~~has~~ several inliers from exit zone of C_3, so ~~adding~~incorporating elements of $\mathbf{Z}$ is ~~same as~~equivalent to ~~giving~~enhancing ~~significant~~the magnitude of the global minimum ~~to matrix P~~in the optimization. Thus, the incorrect~~wrong~~ exit/entry pairs correspondences decrease as the percentage of the outliers in the training data diminished. As a result, the resulting camera link models estimated becomes more accurate with utilizing more true positive pairs and less false positive pairs.

3.3. Bidirectional Link Model

In [3], the camera link model is constructed based only on ~~just~~ one exit/entry direction (e.g., exit from C_1 and entry on C_2) and that was applied to both directions (of link1, i.e., exit from C_1 and entry on C_2 as well as exit from C_2 and entry on C_1) to find correspondence. The camera link model contains several components: transition time distribution,

Jenq-Neng Hwang
Formatted: Font: Italic

Jenq-Neng Hwang
Formatted: Normal

Jenq-Neng Hwang
Formatted: Font: 10 pt

Jenq-Neng Hwang
Formatted: Normal, Cent

Jenq-Neng Hwang
Formatted: Font: 10 pt

Jenq-Neng Hwang
Formatted: Font: 12 pt

Jenq-Neng Hwang
Formatted: Normal

Jenq-Neng Hwang
Formatted: Font: Bold

Jenq-Neng Hwang
Formatted: Font: Italic

Jenq-Neng Hwang
Formatted: Font: Italic, Subscript

Jenq-Neng Hwang
Formatted: Font: Italic

Jenq-Neng Hwang
Formatted: Font: Italic, Subscript

Jenq-Neng Hwang
Formatted: Font: Italic

Jenq-Neng Hwang
Formatted: Font: Italic, Subscript

Jenq-Neng Hwang
Formatted: Font: Italic

Jenq-Neng Hwang
Formatted: Font: Italic, Subscript

Jenq-Neng Hwang
Formatted: Justified

Jenq-Neng Hwang
Formatted: Font: Italic

Jenq-Neng Hwang
Formatted: Font: Italic, Subscript

Jenq-Neng Hwang
Formatted: Font: Not Bold

Jenq-Neng Hwang
Formatted: Font: Italic

Jenq-Neng Hwang
Formatted: Font: Italic, Subscript

Jenq-Neng Hwang
Formatted: Font: Italic

Jenq-Neng Hwang
Formatted: Font: Italic, Subscript

Responses to Reviewers for TCSVT 7047 Paper Entitled: "Fully Unsupervised Learning of Camera Link Models for Tracking Humans Across Non-overlapping Cameras"

Dear Dr. Zhang and Reviewers,

We would like to express our gratitude for your great review efforts and the valuable comments, which have substantially helped in improving the revision. This response letter includes amendments that have been incorporated in the revision and also the responses to the reviewers' comments. We have tried our best to comply with the Reviewers' comments as much as possible. The bold and black texts are the comments from reviewers, and the blue plain texts are the replies to the corresponding comments.

Best Regards,
The Authors

Reviewer 1:

The authors present a multiple-camera tracking systems for non-overlapping cameras. The main contribution lies in a camera link model and the methods to estimate the model. The system is tested on two small datasets and demonstrated good results. In general, the proposed system seems practically sound for small scale camera networks, as shown in the results. In addition, the paper is very well organized and written, and hence easy to follow.

Reply: We appreciate the reviewer for the comments.

A major concern is that, while the authors imply that the proposed system beats previous ones in its scalability (page 1, right bottom para, "as the scale of the camera network is getting larger ..."), the experiments lack in this aspect. The two datasets involve only four and two cameras, respectively, which could be easily handled by supervised systems. Larger scale experiments are imperative to fully evaluate the proposed system.

Reply: We think it is fair to claim the advantage of unsupervised learning over supervised learning when speaking of scalability. In supervised learning, whenever the cameras are set up, people need to manually identify the correspondence in the training data. When more cameras are added into the network, huge amount of human efforts are required. In unsupervised learning, the system automatically learns the camera link models. Moreover, the learning is in pairwise manner, so each link is learned independently. Thus, it is reasonable to expect unsupervised learning is more feasible and scalable in large scale of network.

However, we also understand the concern about the limited datasets used in the paper. Currently, it is not easy to collect data from any large-scale camera networks. Due to the privacy invasion issues, we had to fight greatly with the University administration to install 4 cameras surrounding our EE building with many restrictions imposed by the school. To tone down our claims, we have revised our paragraph in page 1, by saying that the proposed unsupervised learning can help to reduce tremendous amount of manual work while obtaining reasonable results. This is the main idea behind the potential solution for scalability issue encountered in deployment of larger scale camera networks. As shown in the estimation results in Section VII.A that our estimation matches well the one from supervised learning, which requires manually labelled ground truth information. Only through the unsupervised learning of camera link models, which can be continuously updated without further provision of manually labelled ground truth for changing scenarios, the consistent tracking across camera networks can be made feasible.

Another concern is that there are quite some parameters need to be tuned in practice. This may bring the generalization problem, especially considering the lack of enough experiments discussed above.

Reply: Since it is unsupervised learning, modeling the problem with some parameters enables us to get better solution. We admit there are some parameters tuning work need to be done before applying our proposed method. Among the parameters, only two of them are critical to our training stage: β and θ. We have honestly discussed the influence of these two parameters in Section VII.C. The other parameters, like the number of iterations, σ_{true} in (11), etc, are not as important as those two.

In Eq. (19), when using entropy we need to assume that P_ij are elements of probability distributions. This may not be true for the N_1+1 row and N_2+1 column.

Reply: We agree that it is not true for the N_1+1 row and N_2+1 column. Actually, Eq. (19) comes from considering rows and columns separately, and then we made some simplification without affecting the results.

According to the constraint Eq. (23), $\sum_{i=1}^{N_1+1} P_{ij} = 1 \ \forall j \leq N_2$, $\sum_{j=1}^{N_2+1} P_{ij} = 1 \ \forall i \leq N_1$, the entropy cost function should be the combination of the entropy functions for row and column separately:

$$cost_{entropy} = \frac{1}{\beta}(\sum_{i=1}^{N_1}\sum_{j=1}^{N_2+1} P_{ij}logP_{ij} + \sum_{i=1}^{N_1+1}\sum_{j=1}^{N_2} P_{ij}logP_{ij}).$$

We can rearrange it into $cost_{entropy} = \frac{1}{\beta}(\sum_{i=1}^{N_1}\sum_{j=1}^{N_2+1} P_{ij}logP_{ij} + \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} P_{ij}logP_{ij})$. (As stated in the paper, all the discussion should automatically exclude P_{N_1+1,N_2+1}). If we use this as the new cost function, the update equation for P matrix becomes

$$\frac{\partial J(P)}{\partial P_{ij}} = 0 \ \rightarrow \ P_{ij} = \begin{cases} e^{\frac{\beta}{\sigma_{true}}(\theta - dist_{ij})-1)} & if \ i \leq N_1 \ and \ j \leq N_2 \\ e^{-1} & if \ i = N_1+1 \ or \ j = N_2+1 \end{cases}.$$

Compared to Eq. (28), they differ only in the coefficient of the term $(\theta - dist_{ij})$ in exponential function, i.e., the β. If we choose the range of β properly, they are actually the same. Therefore, we simplify the entropy cost function by discarding the second term $\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} P_{ij}logP_{ij}$, and the entropy cost becomes $cost_{entropy} = \frac{1}{\beta}\sum_{i=1}^{N_1+1}\sum_{j=1}^{N_2+1} P_{ij}logP_{ij}$

Reviewer 2:

This paper proposes a multi-camera tracking system that tracks humans across cameras with no-overlapping views. The estimation of camera link model is formulated as an optimization problem in which some temporal and appearance features are jointly learned via an unsupervised learning scheme. The paper is overall written clearly but there are too many symbols which make it seem chaos.

Reply: We are sorry for using many symbols, which are inevitable due to a quite complicated constrained optimization formulation that involves many contributing terms. We use symbols so that it is easier for readers to verify the correctness of technical part and also to generalize it in the future. We did try our best to remind the readers about the meaning of those symbols whenever possible so that they will not get confused.

What is the difference between the optimization formulation (21) and formulation (2) in [22]? Related explanations should be given.

Reply: Eq. (2) in [22] includes linear square as the cost function while we design our own cost functions, $cost_{time}$, $cost_{holistic\_color}$, $cost_{region\_color}$, and $cost_{region\_texture}$. Moreover, due to the characteristics of the features in our problem, we include constraints (24)-(26) in our work. The entropy cost $cost_{entropy}$ and outlier cost $cost_{outlier}$ are the same as the one in [22].

We would like to emphasize that our contribution is to apply the technique used in image points matching problem to the camera link model estimation problem. To the best of our knowledge, we are the first one solving this problem by using this fully unsupervised method. Moreover, it is not a trivial work to apply the technique. We have designed our own cost functions and constraints based on the features we used, and we also need to come up with the iterative optimization approach to effectively solve the problem.

Three different features (i.e. holistic color feature, region feature and texture feature) are combined. What is the effect of each feature? Related experiment evaluation should be given.

Reply: Actually there are four features: temporal, holistic color, region color, region texture. We have added the effect of each feature and some combination between them into Table IV. Readers can now see the individual influence of using these features. Some explanation is also included in the Section VII.B.

When we use only a single feature (either temporal, holistic color, or region color), to perform the tracking, they all achieved roughly around 60% accuracy, while using only region texture feature can achieve worse performance than those three. Since the sizes of the objects are not large enough in some video clips, the texture feature sometimes cannot characterize the object well. If we use multiple features with adaptive fusion weights, the accuracy increases because different features compensate the deficiency of each other. For instance, when temporal and holistic color features are considered (feature set 1,2), similar to the methods in [12] and [20], the re-identification accuracy is 69%. If we combine temporal feature with region features (feature set 1,3,4), the accuracy further improves. However, if we consider multiple features without adaptive fusion weights, the accuracy does not necessarily improve. For example, the result of feature set 1,2,3,4 with uniform fusion weight is a little bit worse than the one of feature set 1,3,4. By further incorporating the adaptive feature fusion weights, our system achieves 79.5% accuracy. Compared to the existing multiple-camera tracking system [12][20], our proposed system enhances the performance by considering region matching and feature fusion weights.

In Fig.3, do the distributions of positive and negative samples really satisfy the Gaussian distribution? Related experiments should be given.

Reply: In the previous version of the manuscript, we did not claim that they satisfy Gaussian distribution since there is no way to prove whether they are drawn from Gaussian distribution or not. Figure 3 was just a plot for illustrating how the d-prime metric measures the degree of separation given two distributions. We are sorry to make it misleading. In current version, we have replaced Figure 3 with another figure which includes the real distributions and also the fitted Gaussian curves built based on the mean and variance. Although they do not perfectly match, the separation between two real distributions are similar to it between two Gaussians, so we think the d-prime metric can be employed for the measurement of degree of separation.

How to derive the update formulations (44) and (45)? If based on maximum likelihood estimation method, the two formulas may have other forms as proposed by the following paper: Zhang, Kaihua, and Huihui Song. "Real-time visual tracking via online weighted multiple instance learning." Pattern Recognition (2013).

Reply: Since the model update scheme is not our major focus, we applied conventional exponentially weighted moving average and variance as our update scheme. Given a newly generated data, it is combined with previous model based on learning rate in order to update new model. Hence, it does not discard the existing model but gradually update the model with new data. This method is also used in background modeling: C. Stauffer and W.E.L. Grimson, "Adaptive background mixture models for real-time tracking", *CVPR*, 1999.
We appreciate the reviewer to point out the paper with more sophisticated method for updating model parameters. We have cited and included this paper into the reference list so as to provide the readers an alternative for model update.

The experiment evaluations are insufficient. More state-of-the-art algorithms should be compared and more data sets should be used. Moreover, some criterions should be used to verify the effectiveness of the proposed method.

Reply: Our future work is to enhance the scalability of the proposed system and also deploy the proposed system in a larger scale camera network. At this point, we are planning to set up more cameras in proper locations (a testbed site is being constructed in a campus outside US since we have some hard time to convince UW administration for larger scale camera installation due to the privacy invasion issues), but we do not have more data set available yet. However, we have tried our best to provide explanation and additional experiments based on the reviewer's comments above. We hope they are sufficient for the reviewer to know more details about our proposed methods.

To the best of our knowledge, there are not many works focusing on fully unsupervised camera link model estimation. The comparison methods [15~18] are indeed state-of-the-art algorithms for this problem, and all of them are published in well-known computer vision journals and conferences. We believe that it is reasonable and meaningful to compare with them.

We are not sure about the criterions the reviewer mentioned. We compare our transition time distribution with the ground truth and competing methods and provide quantitative results. Also, the re-identification accuracy is presented for overall tracking system. These two evaluation metrics are quite popular in this topic. We think they indeed show the effectiveness of the proposed method.

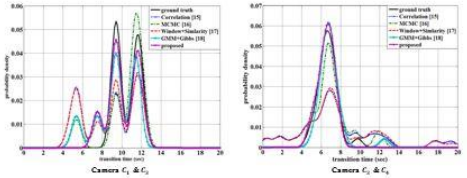
Reviewer 3:

Some important references are missing.
[1] Free Activity Analysis and Scene Modeling in Multiple Camera Views. X. Wang, K. Tieu, and E. Grimson. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 32, pp. 56-71, 2010
[2] Cheng-Hao Kuo, Chang Huang, and Ram Nevatia. "Inter-Camera Association of Multi-Target Tracks by On-line Learned Appearance Affinity Models." European Conference on Computer Vision (ECCV), Crete, Greece, September 2010.

Reply: We thank the reviewer for providing the references. The second reference actually was already in the reference list ([21]) of the previous version. We have included the first one in the reference lists of the current version.

In table I, the transition errors of Camera 2 & 3, Camera 1 & 2 are presented. What about the results of Camera 1&3 and Camera 3&4?

Reply: The distributions between Camera 1&3 and Camera 3&4 (shown below) are relatively simple compared to the other two. They have clear modes in distributions. Due to the limited space, we thought it would be better to show more difficult cases in order to demonstrate the capability of our proposed method, so we did not include them in the previous version. We have decided to include those results in the revised version in order to show the complete results to the readers. Please also see the results for camera 1&3 and 3&4 below.



The implementation details of the methods (e.g. the parameters of Correlation, MCMC, Window+Similarity, GMM+Gibbs) should be clarified.

Reply: In the implementation of the compared methods, we have tried our best to follow the descriptions of the original papers. For those parameters without explaining or suggestion in the paper, we tried several values which we think are reasonable and select the one with the best result for fair comparisons. We summarize the implementation details for each method below:

- Correlation [15]: the time search window T defined in the paper is set as 20 and 35 for our data set and iLIDS data set, respectively.
- MCMC [16]: The Metropolis-Hastings algorithm is utilized. Three types of proposal for sampling procedure: add a match, delete a match and flip two matches. The entropy is used when computing likelihood of the sample. It takes 100~250 iterations to converge depending on the size of dataset. Since random sampling scheme is involved, we report the best result among 30 trials.
- Window+Similarity [17]: the reappearance period T (time window) is set the same as above; Histogram intersection is utilized to build the distribution (the same as the original paper).
- GMM+Gibbs [18]: Gibbs sampling are used for finding correspondence. GMM is employed to generate the distribution. We have tried different penalty parameter α from 0.1 to 2. The number of Gaussians in GMM is set as 3, the same as the original paper. Normally 100 iterations are enough for convergence. Similar to MCMC [16], random sampling scheme is involved, so we report the best result among 30 trials.

What does the word "outlier" mean in this paper? I'm surprised I cannot find its definition through the paper. What' the relationship of "exit observations", "entry observations", "matched pairs" and "outliers" in experiments?

Reply: Actually we indeed mentioned it both in Section I and Section III.A in previous version. We are sorry without making it clear enough, so we have rephrased the paper and explicitly defined it in the paper.

Between a pair of directly connected cameras, exit observations are people who have left one camera's view, and entry observations are people who have entered into the other camera's view within the same period of time window. The matched pairs are a pair of exit and entry observations who are the same person. Outliers are the remaining exit/entry observations who only appear in exactly one camera between this pair of cameras, either an exit observation from one camera never enters the other camera later within the specified time window period or an entry observation to one camera never exit from the other camera earlier within the specified time window period. For instance, if there are four exit observations $x_1 \sim x_4$ in one camera and five entry observations $y_1 \sim y_5$ in the other camera. Assume the matched pairs are (x_1, y_1), (x_2, y_2) and (x_3, y_3). In this case, x_4, y_4 and y_5 are outliers. Therefore, the outlier rate is 3/(4+5) = 33%.

In page 1, right column, line 7, "within a camera's view, in a camera's view" should be "within a camera's view"

Reply: Thank you for the correction. We have changed the sentences to the way that best conveys the idea.

Graduation (One-Day Happiness) and Job Hunting

- Academics (research Labs) or industry?
- As many publications as possible for academic jobs.
- Practical and skilled implementation and broad knowledge for industry jobs.
- Strong presentation and Q&A skills
- Good jobs to push for graduation
- Sell yourself via connections everywhere (technical meeting, industrial affiliated programs, sponsored research review, self-invited talks, etc)
- Sell yourself through a well designed social site (e.g., Linkedin), your skills and achievements

Thanks for your attention!

Best Wishes to Your Graduate Study