REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
29-11-2017	Final Report	1-Aug-2016 - 30-Apr-2017

4. TITLE AND SUBTITLE
Final Report: Multiscale Path Metrics for the Analysis of Discrete Geometric Structures

5a. CONTRACT NUMBER
W911NF-16-1-0392

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
Duke University
C/O Office of Research Support
2200 W. Main St., Ste. 710
Durham, NC 27705 -4677

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
69499-CS-II.4

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Carlo Tomasi
UU	UU	UU	UU		19b. TELEPHONE NUMBER 919-660-6539

RPPR Final Report
as of 30-Nov-2017

Agency Code:

Proposal Number: 69499CSII **Agreement Number: W911NF-16-1-0392**
INVESTIGATOR(S):

Name: Carlo Tomasi
Email: tomasi@cs.duke.edu
Phone Number: 9196606539
Principal: Y

Organization: **Duke University**
Address: C/O Office of Research Support, Durham, NC 277054677
Country: USA
DUNS Number: 044387793 EIN: 560532129
Report Date: 31-Jul-2017 Date Received: 29-Nov-2017
Final Report for Period Beginning 01-Aug-2016 and Ending 30-Apr-2017
Title: Multiscale Path Metrics for the Analysis of Discrete Geometric Structures
Begin Performance Period: 01-Aug-2016 **End Performance Period:** 30-Apr-2017
Report Term: 0-Other
Submitted By: Carlo Tomasi Email: tomasi@cs.duke.edu
Phone: (919) 660-6539
Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: 1 **STEM Participants:** 3

Major Goals: The objective of this research is to model the topological and combinatorial structure of realistic data such as video that captures moving objects (pedestrians, vehicles, marine vessels) in the context of recognition, tracking, and classification tasks. Each data item (for instance, an image of a handwritten digit, or the bounding box around the image of an observation target in one video frame) is viewed as a point in an abstract space, and the main thrust of this research is to model the topology, geometry, and combinatorics of the sets that result from collecting many points like these from data recordings.
This effort explored two directions in this framework:
The development of descriptors of data sets based on path metrics, measures of how strongly different data points are connected by chains of local transformations.
Methods for describing data similarity in the context of visual tracking, where the local transformations mentioned above involve non only similarity of appearance, but also closeness in space and time.

Accomplishments: Please see uploaded PDF document.

Training Opportunities: Work under this grant has been developed in collaboration with a graduate student, Ergys Ristani, who will use results from his research as part of his PhD thesis defense.
Francesco Solera, a PhD student at the University of Modena and Reggio Emilia, in Italy, has contributed to this work while a visitor to Duke Computer Science. Dr. Solera has recently graduated with work that includes his contributions to this project.
Shuai Yuan, an undergraduate Computer Science student at Nanjing University, China, has developed some aspects of this work in collaboration with the PI. Shuai is now applying for a PhD in the Computer Science Department at Duke.
Dr. Paolo Emilio Barbano, a Signal Processing Engineer who visited the Mathematics Department at Duke University in 2016, has contributed to several aspects of this work.

Results Dissemination: Work deriving from this research has been published in the IEEE Transactions on Circuits and Systems for Video Technology and at the ECCV Workshop on Benchmarking Multi-Target Tracking, and these publications, which acknowledge ARO support, have been uploaded with this report.

Honors and Awards: While working on this grant, the PI, Carlo Tomasi, was named a Fellow of the Association of Computing Machinery. Duke University also named him the Iris Einheuser Distinguished Professor. Both honors were given in 2016.

RPPR Final Report
as of 30-Nov-2017

Protocol Activity Status:

Technology Transfer: Nothing to Report

PARTICIPANTS:

Participant Type: Graduate Student (research assistant)
Participant: Ergys Ristani
Person Months Worked: 6.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Francesco Solera
Person Months Worked: 6.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Undergraduate Student
Participant: Shuai Yuan
Person Months Worked: 6.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

CONFERENCE PAPERS:

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: ECCV Workshop on Benchmarking Multi-Target Tracking
Date Received: 29-Nov-2017 Conference Date: 10-Oct-2016 Date Published: 30-Oct-2016
Conference Location: Amsterdam, The Netherlands
Paper Title: Performance measures and a data set for multi-target, multi-camera tracking
Authors: Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, Carlo Tomasi
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 4-Under Review
Conference Name: IEEE Conference on Computer Vision and Pattern Recognition
Date Received: 29-Nov-2017 Conference Date: 18-Jun-2018 Date Published:
Conference Location: Salt Lake City, Utah
Paper Title: Good appearance features for multi-target multi-camera tracking
Authors: Ergys Ristani, Carlo Tomasi
Acknowledged Federal Support: **Y**

RPPR Final Report
as of 30-Nov-2017

Accomplishments

Accomplishments under this effort are described under the two categories of *path distances* and *similarity measures for visual tracking*.

Path Distances

The general trend in describing data collections in current literature is to estimate some "underlying manifold." The existence of such a manifold is questionable for complex data such as images or video. Instead, this project starts from the observation that the transformations between nearby data points are often well understood and domain dependent. For instance, in video, these transformations stem from variations in shading because of small changes in lighting between one frame and the next; small deformations of deformable objects; small variations of appearance because of small camera motions.

The modeling effort then goes into developing models for these transformations. Given a data cloud, one collects statistics $p(T|\text{data})$ of the transformations T_i^j between nearby points i and j, and allowed local transformations are those with high likelihood according to these statistics. A *local distance* is a suitable norm $\|T_i^j\|$ of a local transformation.

Given two dissimilar points, a *global distance* between them can be computed by considering all chains of local transformations that connect them. One such chain is called a *path*. The most likely path can be computed by dynamic programming if the cost of the path is an associative combination of the local costs (examples: sum, max):

$$\hat{\pi} = \arg \max_{\pi \in \text{Paths}} \prod_{T \in \pi} p(T|\text{data}) \ .$$

The distance between the two dissimilar points is some statistical summary of the costs of the local transformations on the most likely path:

$$\mathbb{E}\left[f\left(\bigcup_{T \in \hat{\pi}} p(T|\text{data}) \right) \right] \ .$$

Such a summary is called a *path distance*.

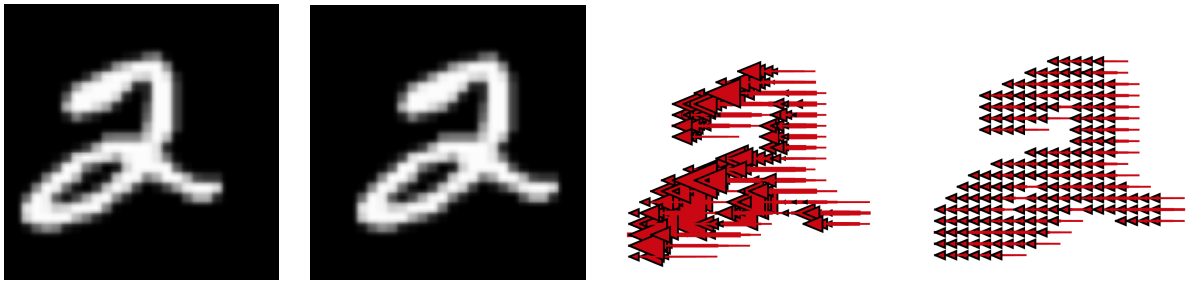
In this way, the underlying structure of the data set can be described in terms of path distances rather than manifolds. Local distances are justified by domain-dependent considerations, and global distances rely on the tight spatial and temporal spacing between observations in dense data streams such as video.

Most of the work on path distances under the short duration of this grant was devoted to the understanding of local distances in the domain of images of hand-written digits from the MNIST data set [2].

Specifically, local distances for hand-written digits were developed as variants of the Earth Mover's Distance (EMD, [5]). Given two images, the EMD first solves a linear program to compute a suitably constrained flow between the mass distributions represented by the two images, and then measures the work (distance times mass) involved in the flow.

For hand-written digits, images are first transformed from Cartesian to log-polar coordinates to achieve rotation and scale invariance.

With the standard formulation of the EMD, the linear program that computes the flow tries to move most mass as little as possible, and this leaves several points that must undergo a large shift. The resulting flow is very unnatural (Figure 1 (c)). If the cost charged to moving a unit mass is made proportional to a power (> 1) of the distance, the flow becomes spatially uniform (Figure 1 (d)). This leads to more natural and more intuitive distances between digit images.



(a) (b) (c) (d)

Figure 1: *(a, b)* Two images of the same hand-written digit that differ by a horizontal shift. *(c)* The Earth Mover's Distance charges cost md to move a mass m by Euclidean distance d. The flow that transforms one image to the other with minimal work is very unnatural. The size of an arrow tip is proportional to the distance traveled. *(d)* Changing the unit cost from d to d^r where $r = 1 + \epsilon$ with $\epsilon > 0$ yields the "natural" flow.

In particular, the modified distance yields to adequate path distances in the MNIST data set. This is best illustrated by displaying digits through multi-dimensional scaling [6], which transforms EMD-based distances to Euclidean distances on the plane. While samples of the digits 0 and 1 appear as jumbled together with the regular EMD, Figure 2 shows that the new variant, which includes transformation to log-polar coordinates and a power $r = 1.25$, gives good separation between the two groups of digits (with a few exceptions) and this improvement is likely to lead to better recognition and classification algorithms.

Similarity Measures for Visual Tracking

When tracking people in video, it is necessary to determine the similarity or dissimilarity between different images of people, in order to determine if the two images are of the same person or not. This application was used as an opportunity to study the extent to which local dissimilarities can be *learned*, rather than hand-crafted.

To this end, a large collection of snapshots of people [4], developed by the PI under a previous ARO grant (W911NF-10-1-0387), was used as a training set, and a so-called *triplet loss*

$$L = \log(1 + \exp(1 + d(a, p) - d(a, n)))$$

was used to learn the local dissimilarity. In this expression, d is the dissimilarity to be learned, a is a person snapshot (the "anchor" of the triplet), p is a snapshot of the same person as a, and n is a snapshot of someone else. The idea of using a triplet loss is that $d(a, p)$ should be consistently smaller, by a margin of 1, than $d(a, n)$ over all triples of this format.

The proposed approach leverages the idea of PK batches [1] for training, where in each batch there are K sample images for each of P identities. During a training epoch, each identity is selected in its batch, and the remaining $P - 1$ batch identities are sampled at random. K samples for each identity are then also selected at random. The *batch-hard* loss [1] can be defined as:

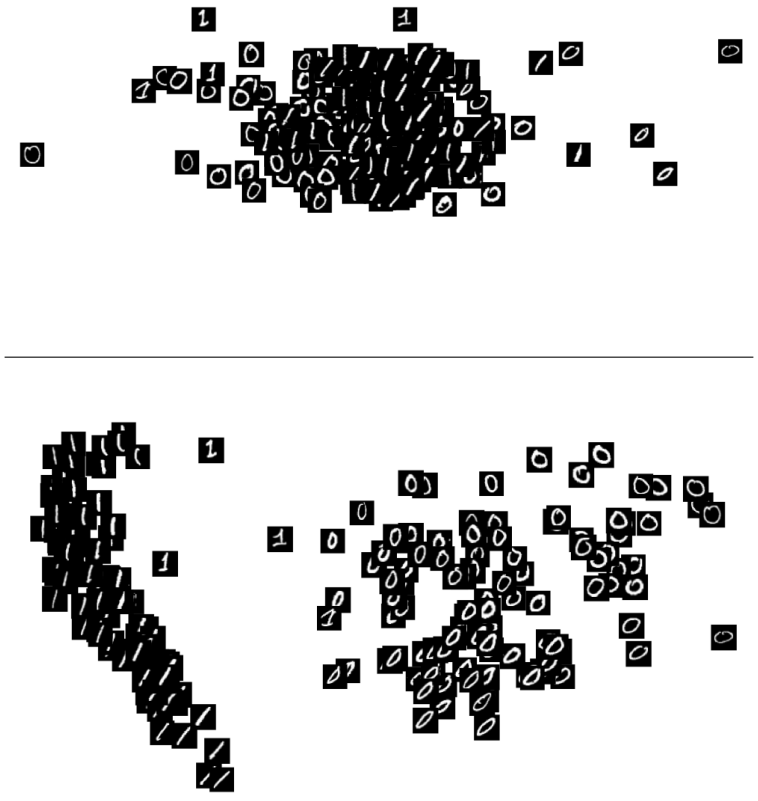


Figure 2: *Top:* The standard EMD jumbles ones and zeros together in this rendering through multidimensional scaling. *Bottom:* The new variant separates the two digits cleanly, with only a couple of exceptions.

$$L = \overbrace{\sum_{i=1}^{P}\sum_{a=1}^{K}}^{\text{all anchors}} \Big[1 + \overbrace{\sum_{\substack{p=1 \\ p\neq a}}^{K} w^{+}_{i,a,p} d(x^i_a, x^i_p)}^{\text{positive pairs}} \underbrace{- \sum_{\substack{j=1 \\ j\neq i}}^{P}\sum_{n=1}^{K} w^{-}_{i,a,j,n} d(x^i_a, x^j_n)}_{\text{negative pairs}} \Big]_+ \tag{1}$$

where the weights are binary:

$$w^{+}_{i,a,p} = \begin{cases} 1, & p = \arg\max\limits_{\substack{k=1\ldots K \\ k\neq a}} d(x^i_a, x^i_k) \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

$$w^{-}_{i,a,j,n} = \begin{cases} 1, & (j,n) = \arg\min\limits_{\substack{l=1\ldots P \\ l\neq i \\ k=1\ldots K}} d(x^i_a, x^l_k) \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

d denotes Euclidean distance, and $[x]_+ = ln(1 + e^x)$.

Training with the batch-hard loss on a PK batch is equivalent to training with a triplet loss on a batch with PK triplets of the structure (x_a, x_{hp}, x_{hn}) corresponding to the anchor x_a, hardest positive example x_{hp} and hardest negative example x_{hn} within the batch.

The main improvement made under this grant is on the procedure that selects difficult examples. As the size of the training set increases, sampling $P - 1$ identities at random rarely picks the hardest negatives, keeping batch difficulty to semi-hard. To increase the chances of seeing hard negatives, two sets are constructed to sample identities from. The hard identity pool consists of the H most difficult identities given the anchor, and the random identity pool consists of the remaining identities. Then in a PK batch of an anchor identity, the remaining $P-1$ identities are sampled from the hard or random identity pool with equal probability. This technique samples hard negatives more frequently and yet the batch partially preserves dataset statistics by drawing random identities.

Equations 2-3 assign full weight to the hardest positive/negative example for each anchor while ignoring the remaining $K - 2$ positives and $K(P - 1) - 1$ negatives. This makes the optimization non-smooth for challenging/mislabeled datasets or very difficult batches. A second improvement was to define weights using a softmax/min as follows:

$$w^{+}_{i,a,p} = \frac{e^{d(x^i_a, x^i_p)}}{\sum\limits_{\substack{k=1 \\ k\neq a}}^{K} e^{d(x^i_a, x^i_k)}}, \quad w^{-}_{i,a,j,n} = \frac{e^{-d(x^i_a, x^j_n)}}{\sum\limits_{\substack{l=1 \\ l\neq i}}^{P}\sum\limits_{k=1}^{K} e^{-d(x^i_a, x^l_k)}} \tag{4}$$

The hardest positive/negative samples still dominate the loss, while all other samples of a PK batch contribute. This modification makes training stable with a minor drop in rank accuracy.

The TriNet deep neural net of [1] is used for learning appearance features. It consists of a ResNet50 model pre-trained on ImageNet whose *pool5* layer is followed by a dense layer with 1024 units and ReLU activation. Batch normalization and another dense layer follow, embedding the appearance features in $\mathbb{R}^d$ with $d = 128$.

These dissimilarities are evaluated experimentally by computing the F1 score (harmonic mean of precision and recall), as well as precision and recall figures obtained on the Duke MTMC data set [4] with a recent method called BIPCC [3]. See Table 1. Applications of this and related ideas were presented in two recent publications [4, 7] and in work under review.

	F1 Score	Precision	Recall
BIPCC	54.98	62.67	48.97
Proposed	**80.26**	**83.50**	**77.25**

Table 1: The proposed dissimilarity achieves much higher F1 score, precision, and recall than BIPCC, a recent state-of-the-art re-identification method [3].

References

[1] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[2] Y. LeCun, L. Jackel, L. Bottou, A. Brunot, C. Cortes, J. Denker, H. Drucker, I. Guyon, U. Muller, E. Sackinger, P. Simard, and V. Vapnik. Comparison of learning algorithms for handwritten digit recognition. In *International Conference on Artificial Neural Networks*, pages 53–60, 1995.

[3] A. Maksai, X. Wang, F. Fleuret, and P. Fua. Non-markovian globally consistent multi-object tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[4] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV Workshop on Benchmarking Multi-Target Tracking*, volume 9914, pages 17–35. Springer LNCS, October 2016.

[5] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, November 2000.

[6] R. N. Shepard. The analysis of proximities: Multidimensional scaling with an unknown distance function, i and ii. *Psychometrika*, 27:125–140,219–246, 1962.

[7] F. Solera, S. Calderara, E. Ristani, C. Tomasi, and R. Cucchiara. Tracking social groups within and across cameras. *IEEE Transactions on Circuits and Systems*, page to appear, 2016.