DATA CLUSTER—©SHUTTERSTOCK.COM/ASTIBUAG, BIG DATA—©SHUTTERSTOCK.COM/13FTSTUDIO

The Role of Visual Assessment of Clusters for Big Data Analysis

From Real-World Internet of Things

by Marimuthu Palaniswami,
Aravinda S. Rao,
Dheeraj Kumar,
Punit Rathore, and
Sutharshan Rajasegarar

Digital Object Identifier 10.1109/MSMC.2019.2961160
Date of current version: 21 October 2020

The Internet of Things (IoT) is playing a vital role in shaping today's technological world, including our daily lives. By 2025, the number of connected devices due to the IoT is estimated to surpass a whopping 75 billion. It is a challenging task to discover, integrate, and interpret processed big data from such ubiquitously available heterogeneous and actively natural resources and devices. Cluster

2333-942X/20©2020IEEE

analysis of IoT-generated big data is essential for the meaningful interpretation of such complex data. However, we often have very limited knowledge of the number of clusters actually present in the given data. The problem of finding whether clusters are present even before applying clustering algorithms is termed the *assessment of clustering tendency*. In this article, we present a set of useful visual assessment of cluster tendency (VAT) tools and techniques developed with major contributions from James C. Bezdek. The article further highlights how these techniques are advancing the IoT through large-scale IoT implementations.

Today, the term *big data* (or, alternatively, *very large data sets*) is used to characterize the exponential increase of structured and unstructured data.

Origin of Big Data

IoT technology has significantly changed the way we live today. Physical objects (or devices) with the ability to sense, process, and communicate information to other devices have enabled the IoT to empower "things" (or devices) to not only connect with other devices but to also control devices in far-off parts of the world [1]. This notion of sensing, processing, communicating, and actuating provides a range of unprecedented opportunities to address the many challenges facing our world.

With the beginning of the 21st century, the evolution of the IoT and the number of devices being used on the Internet is exponential; we see billions of devices from every corner of the world now being connected to the Internet. For example, in 2018, there were roughly 23 billion Internet-connected devices when compared with approximately 15 billion in 2015. By 2025, the number of IoT-connected devices is predicted to surpass a whopping 75 billion, a fivefold increase in 10 years [2]. We see IoT penetration in nearly every major industry sector, including agriculture and food, health care, energy and natural resources, water management, transportation and logistics, manufacturing, retail and advertisement, government, insurance, and education. Specifically, we see high penetration in application areas related to precision farming, wearable devices, smart homes, connected vehicles, industrial robots, and smart cities. These IoT devices are producing a mind-boggling, 2.5 quintillion B (i.e., 2.5×10^{18}) of data everyday (in 2018) [3]. The enormous amount of data generated by devices poses several challenges to acquiring, storing, processing, visualizing, and interpreting the data and using that knowledge to our own benefit.

In 1997, Michael Cox and David Ellsworth from NASA first used the term *big data* for data sets that are too large to visualize [4]. In 2001, Doug Laney identified the three Vs (volume, velocity, and variety) of data growth for understanding and dealing with big data [5], [6]. Today, the term *big data* (or, alternatively, *very large data sets* [7]) is used to characterize the exponential increase of structured and unstructured data, which pose challenges to capturing, storing, managing, and processing using conventional data management and analysis techniques. It is a difficult task to discover, access, process, integrate, and interpret data from such ubiquitously available heterogeneous and actively natural resources and devices [8]. Since 2001, the number of "V"s used to identify the different characteristics of big data has grown to seven: volume, velocity, variety, validity, volatility, variability, and visualization [9].

The Visual Assessment of Clusters

Clustering, or cluster analysis, plays a major role in identifying the clusters, i.e., patterns, of subsets' data and relates to the problem of separating a set of data objects $O = \{o_1, o_2, \ldots, o_n\}$ into c self-similar subsets. These subsets are formed depending on the available data and some explicit measure of the similarity of clusters [10]. Depending on the data, geometric descriptions of the clusters are also sought. Although clustering was conceived as an act of segregating the objects into convenient groups, cluster analysis aims to answer [11] 1) how many clusters are there? (the tendency of clusters); 2) to which cluster do the objects belong and to what degree are they associated? (the data partition); and 3) are the partitions of data are good? (validity of the cluster). Clustering algorithms require the number of clusters as input; however, many times we may not know this beforehand, and sometimes it is not possible to determine it by looking at the data. This is where the VAT algorithm [10] comes into picture. It is important to note that, even if no "actual" clusters exist in the data, all of the clustering algorithms will be able to find c number of clusters, where $1 \leq c \leq n, n \in \mathbb{N}$. As a result, it is fundamentally important to ask oneself whether there are any "actual" clusters before applying any clustering algorithms [10].

VAT Technique

The problem of finding whether clusters are present even before applying clustering algorithms is termed the *assessment of clustering tendency*. Several techniques, both formal (based on statistics) and informal (other approaches), have been proposed [12], [13], but they are not completely effective. On the other hand, visual approaches [14], [15] for analyzing data have been around for more than four decades, forming the basis for many visual data-analysis techniques. Bezdek's VAT tool [10] presents pairwise dissimilarity knowledge about the set of objects, $O = \{o_1, o_2, \ldots, o_n\}$. This is usually represented as a square digital image with n^2 pixels

[10]. The advantage of VAT as opposed to other visual techniques is its ability to highlight the potential number of clusters in the data by suitably reordering the objects. This is achieved by reordering the dissimilarity matrix of the input data using a modified Prim's algorithm and visually estimating the number of clusters that appear as the dark blocks along the diagonal of reordered dissimilarity image (RDI). Figure 1 shows how a VAT tool helps determine the number of clusters from a dissimilarity matrix. The VAT tool is widely applicable to large real-world data sets, including big data. It also allows for the display of reordered dissimilarity data, which can be accessed from the original data, O. If O has any missing components, then any existing data-imputation schemes can be used to fill in the missing parts of the data before applying a VAT.

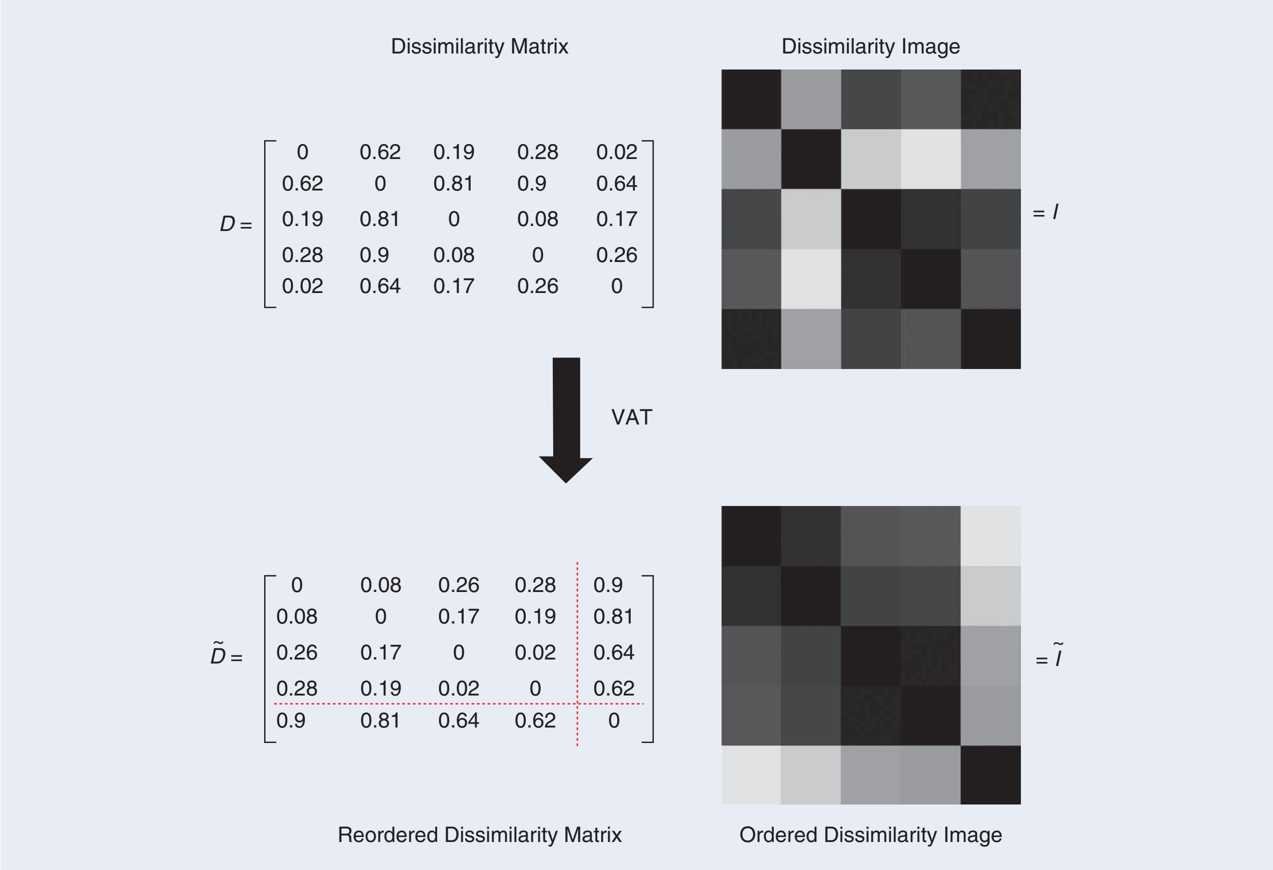
Extensions of VAT for Handling Big Data

An Improved VAT for Improved Contrast

Suppose we have pairwise dissimilarity matrix D of a set of n objects, a VAT generally portrays D as $n \times n$-image $I(\tilde{D})$, where the objects are reordered to reveal hidden cluster structures as "dark blocks" along the diagonal of the image [16]. However, a VAT fails to clearly display the dark block if the data have complex structures. On the other hand, an improved VAT (iVAT) proposes the improved performance of a VAT by transforming the reordered dissimilarity matrix $\tilde{D}$ using a graph-theoretic geodesic distance. Evidently, iVAT significantly enhances the separation of the "dark blocks" in VAT images [11].

A Scalable Visual Assessment of Cluster Tendency for Large Data Sets

Although the VAT tool finds its usefulness in many IoT applications, it can be computationally expensive as the size of the data set grows. An algorithm is said to be scalable if there is a linear increase in the runtime complexity with the increase in the number of observations in the input data [17]. A VAT has a runtime complexity of $O(N^2)$, which is not attractive for large data sets. On the other hand, the scalable visual assessment of cluster tendency (sVAT) algorithm uses a sample-based version of a VAT that can handle large data sets [18]. An sVAT chooses sample size n from the complete set of objects,



$$D = \begin{bmatrix} 0 & 0.62 & 0.19 & 0.28 & 0.02 \\ 0.62 & 0 & 0.81 & 0.9 & 0.64 \\ 0.19 & 0.81 & 0 & 0.08 & 0.17 \\ 0.28 & 0.9 & 0.08 & 0 & 0.26 \\ 0.02 & 0.64 & 0.17 & 0.26 & 0 \end{bmatrix}$$

$$\tilde{D} = \left[\begin{array}{cccc:c} 0 & 0.08 & 0.26 & 0.28 & 0.9 \\ 0.08 & 0 & 0.17 & 0.19 & 0.81 \\ 0.26 & 0.17 & 0 & 0.02 & 0.64 \\ 0.28 & 0.19 & 0.02 & 0 & 0.62 \\ \hdashline 0.9 & 0.81 & 0.64 & 0.62 & 0 \end{array}\right]$$

Figure 1. An illustration of how a VAT reorders the dissimilarity matrix. D is the dissimilarity matrix obtained from objects in O. From D, we note that it is difficult to determine how many clusters are present. I represents the dissimilarity image, $\tilde{D}$ is the VAT reordered matrix obtained after applying VAT, and $\tilde{I}$ is the reordered dissimilarity image. From the pair $(\tilde{D}, \tilde{I})$, we note that there are two clusters of block sizes 4 × 4 and 1 × 1 in $\tilde{D}$, and it is evident from the visual substructure suggested by $\tilde{I}$.

$O=\{o_1, o_2, \ldots, o_N\}$, and executes a VAT on the distance matrix of n sample. The sample is selected such that it contains a cluster structure identical to the full set. This operation requires one to first pick a set of k' distinguished objects using a maximin sampling [19] to provide a depiction of each of the clusters. Subsequently, the remainder of the sample is produced by selecting additional data near each of the distinguished objects, leading to maximin and random sampling (MMRS).

The advantage of VAT as opposed to other visual techniques is its ability to highlight the potential number of clusters in the data by suitably reordering the objects.

To assess the cluster tendency of large volume data sets, we apply the VAT algorithm on the MMRS samples. In Figure 2(a), we see the scatterplot of $N=1{,}000{,}000$ 2D points drawn from four Gaussian components and 250,000 points per cluster. However, we cannot create a VAT image, as shown in Figure 2(b). On the other hand, an sVAT and siVAT allow us to create images by sampling $n=500$ points (0.05% of the total data set) from O. In Figure 2(c), the sVAT image indicates that there are four clusters, which are enhanced in the siVAT image [Figure 2(d)].

A Scalable, Single-Linkage Visual Assessment of Cluster Tendency for High-Volume Data Sets

Although the sVAT is helpful in understanding the cluster structure of the big data and determining the optimal value of k—the number of clusters to seek based on the visual evidence—on their own, they do not partition the data into k clusters. To tackle this issue, the scalable, single-linkage visual assessment of cluster tendency (sVAT-SL) [20]

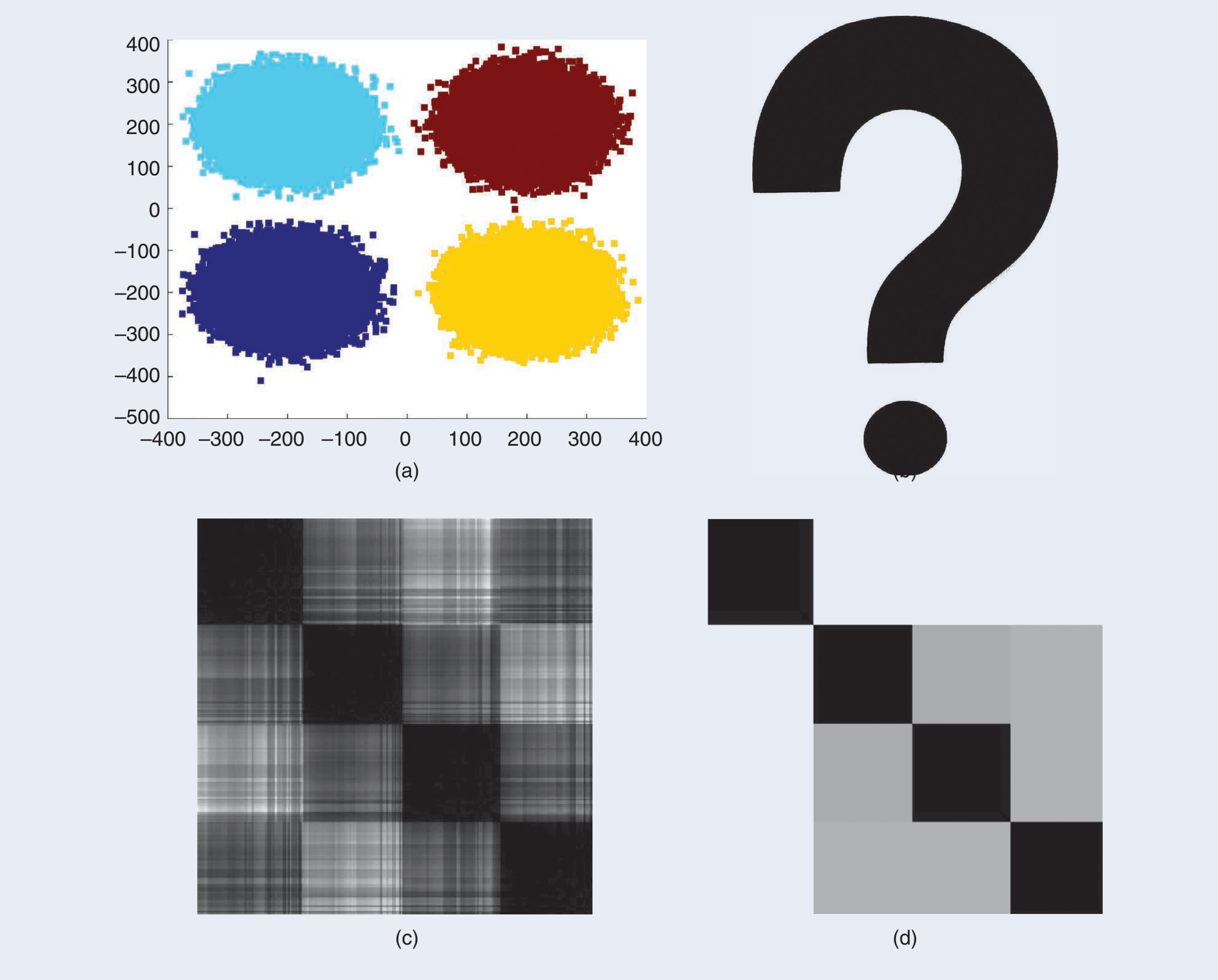


Figure 2. Images of big data Gaussian clusters. (a) The data scatterplot, (b) VAT, (c) sVAT, and (d) siVAT.

extends the sVAT algorithm to return single-linkage partitions of big data. The sVAT-SL works by calculating a single-linkage partition of the sVAT-sampled data and then extending this partition to the entire data set using a nearest prototype rule (NPR). It is shown to be a scalable instantiation of single-linkage clustering for data sets that contain k compact-separated clusters, and the sVAT-SL produces a good approximation of single-linkage partitions on data sets not containing compact-separated clusters [20].

Clustering siVATs and Fast-Clustering siVATs for Faster Computations

The clustering siVAT (clusiVAT) algorithm is superior in regard to cluster quality and computation time over popular big data clustering algorithms, such as minimum spanning trees (MSTs), which are constructed with Filter-Kruskal, k-means, single-pass k-means, online k-means, and clustering using representatives [21], [22]. The clusiVAT is fast in cases where the objects are represented by their feature vectors and the Euclidean distance as a distance measurement between objects. This superior computational speed is because the clusiVAT assumes that the distance function computation is quick and that clustering can be executed as a batch; that is, using matrix operations, we can compute the Euclidean distance of a data point from $M \gg 1$ data points as a single operation. However, this fundamental assumption does not hold for many distance measurements applicable to graphs and time series; there are many distance measurements applicable for problems in different domains that are computationally expensive and can only be computed in a pairwise manner. Fast-clusiVAT addresses this time-consuming distance measurement issue by adopting maximin sampling and NPR while maintaining accuracy [23].

Clustering of Streaming Data

Traditional clustering approaches, which provide a fixed outline of each data point's pledge to every group, may not be suitable for handling streaming data with new clusters or the clubbing of existing concentrated data regions for streaming data sets. Incremental algorithms, such as incremental VAT (inc-VAT), incremental iVAT (inc-iVAT), decremental VAT (dec-VAT), and decremental iVAT (dec-iVAT) provide efficient mechanisms to update the VAT and iVAT RDI in the event a new point is added to, or an existing point is removed from, the current data set.

The time complexity of inc-VAT, dec-VAT, inc-iVAT, and dec-iVAT matches fairly with that of VAT and iVAT, respectively. These sets of algorithms find applications in detecting anomalies as well as in sliding-window-based visual cluster assessments for detecting clusters in streaming data in an online fashion. To illustrate the effectiveness of these algorithms in visualizing the evolving nature of cluster structures and computational efficiency compared to VAT/iVAT, an experiment conducted on a 2D Gaussian mixture (X) of five clusters is shown in Figures 3 and 4. Each cluster in Figure 3 has 1,300, 1,000, 400, 1,600, and 700 data points, respectively. The data points in X are ordered based on cluster membership. As a result, the first 1,300

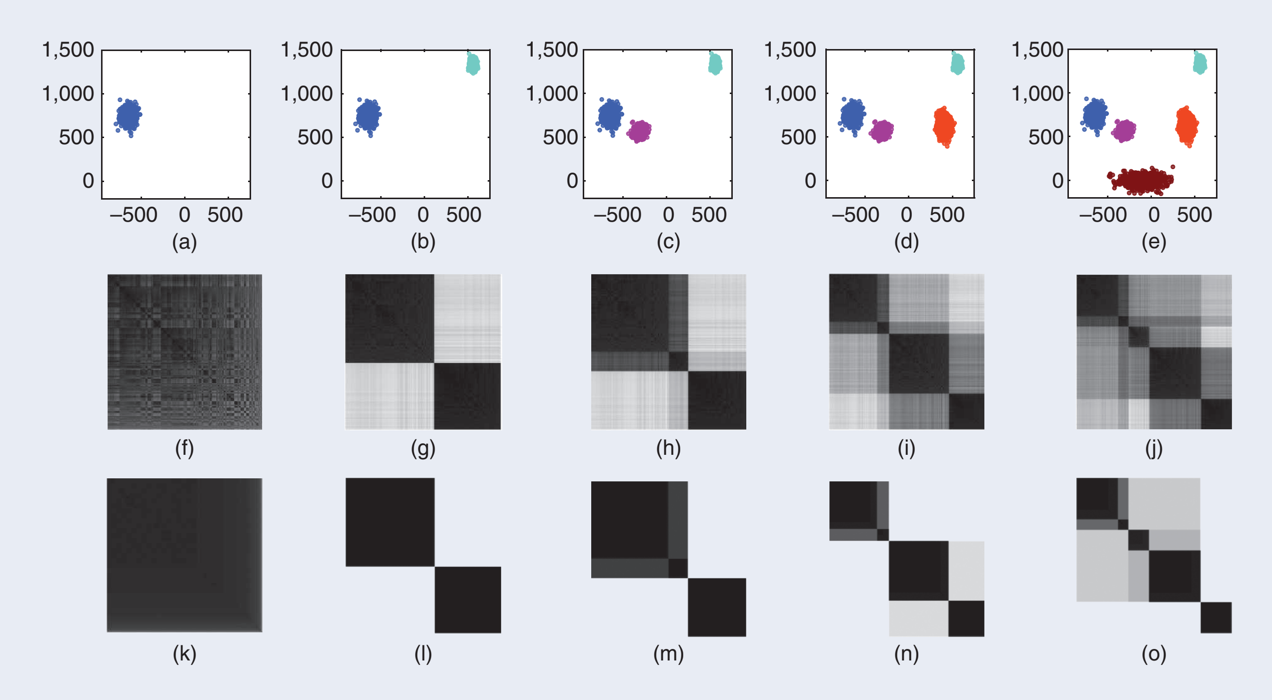


Figure 3. (a)–(e) The 2D data scatterplot, (f)–(j) the inc-VAT, and (k)–(o) the inc-iVAT images of X at $n = 1$, 300, 2,300, 2,700, 4,300, and 5,000.

rows of X are x and y coordinates corresponding to the data points of the first cluster; the subsequent 1,000 rows correspond to the second cluster, and so forth. The different columns of Figure 3 show a subset of X coming from the first cluster, the first two clusters, and the remaining clusters, respectively.

To emphasize the difference in the time complexities of VAT/iVAT and inc-VAT/inc-iVAT, we shuffled the rows of X such that the data points of the same cluster are apart. This was initiated with two data points and then by adding a data point at each time step. At each time step, we measured the time required by each algorithm (VAT, iVAT, inc-VAT, and inc-iVAT) to compute the reordered dissimilarity matrices. From Figure 4(a) we see that, as n increases, VAT + iVAT requires more time to update when compared to inc-VAT + inc-iVAT. Likewise, to reveal the time complexity between dec-VAT/dec-iVAT and VAT/iVAT, we performed the experiment on the aforementioned 2D data (X). Because we were comparing the decremental nature of the algorithms, we initiated the process with $n = 5,000$ data points and eliminated a single arbitrarily selected data point at each time step. From Figure 4(b) we see that, as n decreases, dec-VAT + dec-iVAT requires much less time than VAT + iVAT $[O(n^2)]$.

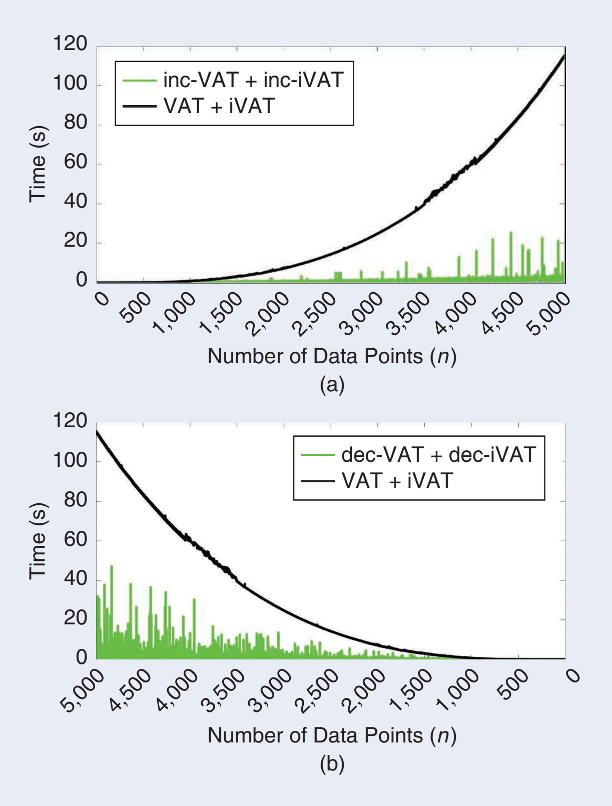


Figure 4. The time required for a combination of VAT, iVAT, inc-VAT, inc-iVAT, dec-VAT, and dec-iVAT algorithms for the 5,000-point 2D data set. (a) VAT + iVAT versus inc-VAT + inc-iVAT and (b) VAT + iVAT versus dec-VAT + dec-iVAT.

Clustering Large Volumes of High-Dimensional Data

The majority of clustering algorithms are designed to handle data sets with either 1) a very large sample size or 2) a very high number of dimensions. However, they are usually impractical when the data set (generated especially from IoT devices) is large (both in sample size and dimensions). From the "Extensions of VAT for Handling Big Data" section, we see that both the sVAT-SL and clusiVAT algorithms have the ability to handle data cardinality with sampling schemes; however, they cannot deal with high-dimensional data. To address this critical issue, Fast clustering by combining ENsemble of random projects with Scalable version of iVAT (FensiVAT) [25] is proposed. FensiVAT is a fast, ensemble-based scalable iVAT algorithm. It integrates a new, random projection-based distance matrix with MMRS sampling and iVAT to cluster large volumes of high-dimensional data. FensiVAT is also several orders of magnitudes faster than alternative clustering techniques, such as clusiVAT, without sacrificing accuracy.

Real-World Applications of VAT families to the IoT

Monitoring the Great Barrier Reef of Australia

The Great Barrier Reef (GBR) of Australia comprises 3,200 coral reefs spanning more than $280,000\ \text{km}^2$ [26]. The GBR is both economically and ecologically sensitive, however, and the burning of fossil fuels has led to the absorption of carbon dioxide in oceans, resulting in acidification of the ocean. This process prevents corals from secreting calcium carbonate exoskeletons, diminishing the reef-building mechanism and its associated organisms. Human-induced activities are increasing the stress on coral reefs, leading to coral bleaching, wherein the symbiotic relationship between the coral and algae breaks down during rapid changes in sea-water temperature (hot or cold) [26].

The Great Barrier Reef Ocean Observing System project aims to provide observational data to determine the long-term effects of the Coral Sea on the ecosystems and the impact on the GBR. To monitor the reef's ecosystem, we collected temperature profiles and weather data from Heron Island in the GBR. The iVAT algorithm detected the passage of Tropical Cyclone Hamish in March 2009 (see Figure 5) [27]. We considered one month of data (21 February–22 March 2009, from 9:00 a.m. to 3:00 p.m., using 10 min of sampling frequency) as a case study. Figure 6 shows the Cyclone Hamish event as two anomalous clusters.

Urban Forest Monitoring in the City of Melbourne

IoT infrastructure for the creation of smart cities consists of Internet-connected sensors, devices, and citizens. This IoT infrastructure generates an enormous amount of data in the form of city-scale physical measurements and public opinions, constituting big data. Smart cities aim to efficiently use this wealth of data to manage and solve the problems

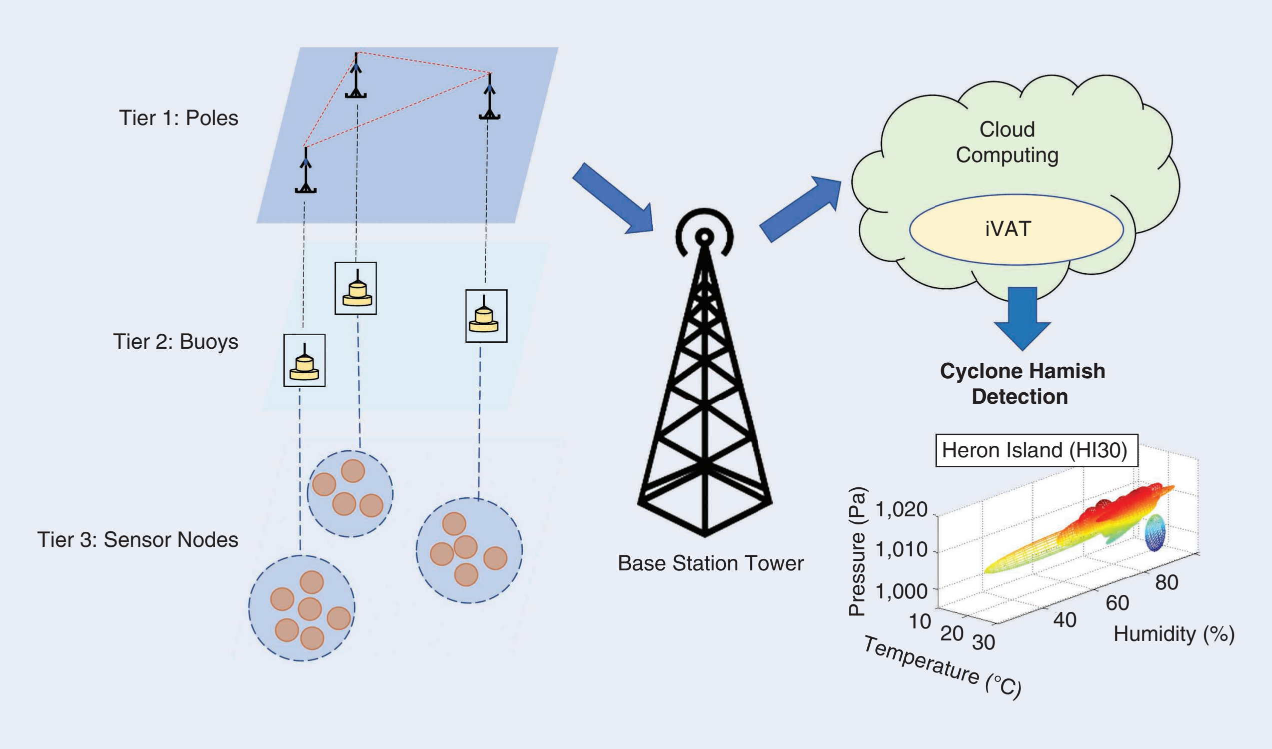


Figure 5. A two-tiered (tree) hierarchical network architecture of wireless sensor nodes deployed on Heron Island of the GBR for the continuous monitoring of the reef. Using iVAT, we detected the passage of Cyclone Hamish (2009).

faced by modern cities for better decision making. However, interpretation of the massive amount of smart city-generated big data to create actionable knowledge is a challenging task. Environmental sensors measuring luminosity, humidity, and temperature were deployed at Fitzroy Gardens and Docklands Library to study the effects of canopy cover and its impact on extrapolating said effects to enhance citizens' interactions with the city's infrastructure.

To this effect, we clustered sensor data (luminosity, humidity, and temperature) from four sensors at the Docklands Library utilizing sliding windows. We resampled the data so that there was one measurement from all four sensors in 30-min intervals, resulting in a 12-dimensional feature vector. In total, we have measurements spanning 72 days; with a window size of two days, we have 96 samples. At each time step, the inc-VAT/inc-iVAT includes a new data point while the dec-VAT/dec-iVAT removes the oldest data point. In the first time step, the inc-VAT/inc-iVAT appends 96 samples to the MST, whereas in the last step, the dec-VAT/dec-iVAT terminates when the sample size equals 2. Figure 7 shows the application of inc-iVAT/dec-iVAT to the Docklands Library data to detect anomalous samples.

Urban Mobility Patterns of Taxis in Singapore

Analyzing clusters is a fundamental challenge in trajectory mining; however, existing trajectory clustering algorithms are not appropriate for large numbers of trajectories in a city's road network because of inadequate distance measurements between two trajectories. We utilized the GPS traces of 15,061 taxis in Singapore (equaling 3.28 million trajectories) gathered over a one-month period. To cluster the origin and destination pairs of taxi rides, we used clusiVAT sampling and the density-based spatial clustering of applications with noise [28] to provide useful insight into urban hot spots, the usage of road networks, and crowd movements. For large numbers of overlapping trajectories, we used dynamic time warping (DTW) coupled with Dijkstra (TrajDTW) distance measurements [23], [29]. For

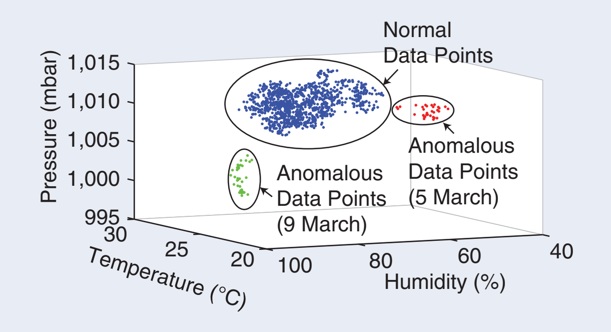


Figure 6. Cyclone Hamish passed the Queensland state of Australia during 4–14 March 2009. The iVAT algorithm accurately clusters these events as two anomalous clusters, corresponding to 5 and 9 March 2009, respectively.

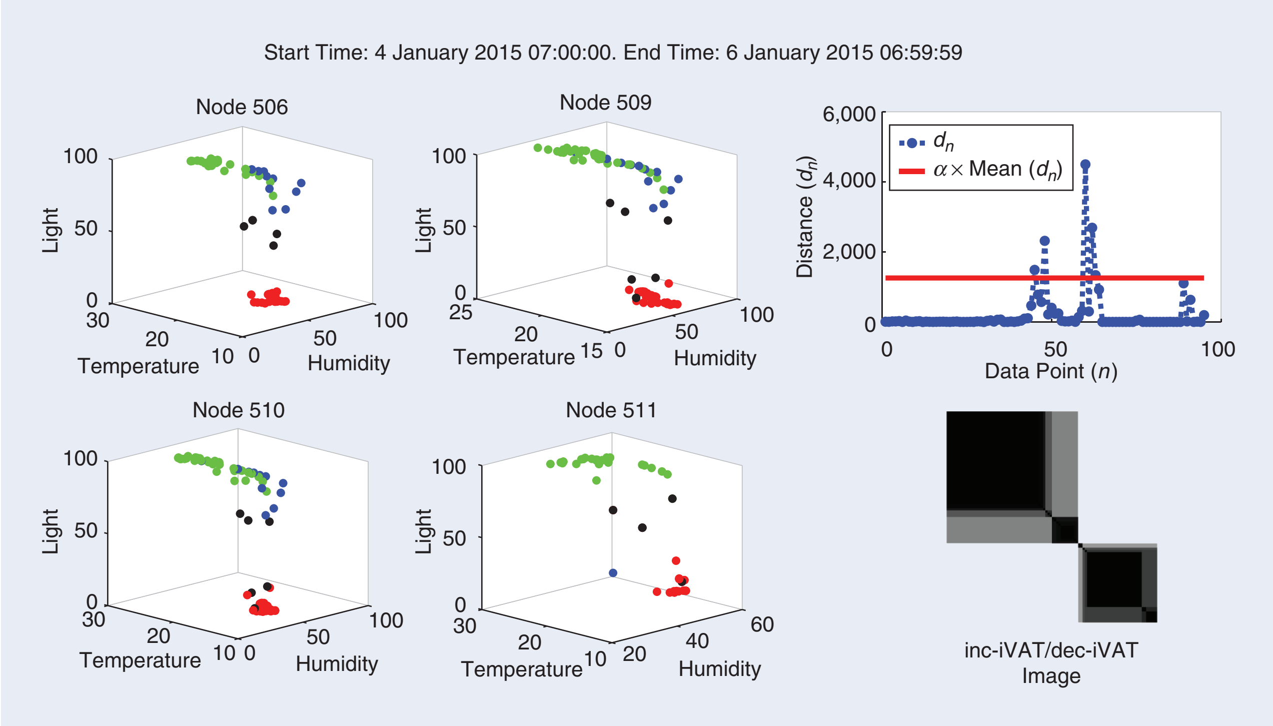


Figure 7. The sensor data of temperature, humidity, and light from each sensor node (510 and 511). In addition, the figure shows the MST-cut magnitude (d_n), and inc-iVAT/dec-iVAT shows the visual estimate of the number of clusters at each time step. For this graph, the parameters used were $\alpha = 6$ and $\beta = 0.1$. The anomalous points are shown using dark dots, whereas the clusters are depicted in red, green, and blue.

predicting trajectories, we utilized Traj-clusiVAT [30], which combines scalable clustering and Markov chains for predicting both short- and long-term trajectories. In addition, Traj-clusiVAT can determine the clusters representing diverse behaviors.

Detecting Anomalies in the Mobility Patterns of Vehicles and Pedestrians

Knowing the templates of pedestrian movement has many useful applications in managing pedestrian flows and maintaining public security and safety. We used iVAT+ and clusiVAT+ [31] for detecting anomalous pedestrian trajectories. These trajectories are classified as normal or abnormal depending on the number of trajectories in the clusters. Experiments on the vehicle and pedestrian trajectories from a parking space data set (https://www.ee.cuhk.edu.hk/~xgwang/MITtrajsingle.html) showcases the ability of a VAT-based approach in producing natural and informative trajectory clusters and finding representative anomalies.

Future Work

The IoT will drive the increased use of connected devices for many applications. This will result in the generation of big data for diverse applications, requiring the analysis of data streams in real time. From the literature and the work presented in this article, we take note of the existing algorithms used for clustering big data, assess the tendency of clusters, and detect the anomalies from big data. However, with the high-velocity streaming data generated by IoT devices, there are very limited algorithms that are 1) suitable for extracting structure from streaming data and that 2) infer the exact number of clusters from the streaming data. Our future work includes designing algorithms that could not only cluster streaming big data but also validate the use of clusters to handle streaming big data.

Conclusion

In this article, we presented an overview of how the family of VAT techniques can be elegantly used to analyze the number of clusters present in the big data generated by IoT devices, even before we apply clustering algorithms. The article explored how Bezdek's pioneering algorithms are effective in analyzing IoT-generated big data. We detailed four real-world IoT case studies (e.g., monitoring impacts on the GBR, monitoring an urban forest, understanding urban mobility patterns, and detecting anomalies in vehicles and pedestrians) wherein VAT techniques and their extensions were applied for solving key issues. These techniques, primarily developed with Bezdek, are advancing the IoT for effective practical implementations.

About the Authors

Marimuthu Palaniswami (palani@unimelb.edu.au) is with the Department of Electrical and Electronic Engineering, the University of Melbourne, Australia. He is a Fellow of IEEE.

Aravinda S. Rao (aravinda.rao@unimelb.edu.au) is with the Department of Electrical and Electronic Engineering, the University of Melbourne, Australia. He is a Member of IEEE.

Dheeraj Kumar (dheerajfec@iitr.ac.in) is with the Department of Electronics and Communication Engineering, Indian Institute of Technology Roorkee.

Punit Rathore (prathore@mit.edu) is with the Senseable City Lab, Department of Urban Studies and Planning, Massachusetts Institute of Technology. He is a Member of IEEE.

Sutharshan Rajasegarar (sutharshan.rajasegarar@deakin.edu.au) is with the School of Information Technology, Deakin University, Burwood, Australia.

References

[1] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," *Future Gener. Comput. Syst.*, vol. 29, no. 7, pp. 1645–1660, 2013. doi: 10.1016/j.future.2013.01.010.

[2] IHS, "IoT: Number of connected devices worldwide 2012-2025," Statista, Nov. 2016. [Online]. Available: https://www.statista.com/statistics/471264/ iot-number-of-connected-devices-worldwide/

[3] B. Marr, "How much data do we create every day? The mind-blowing stats everyone should read," *Forbes*, May 21, 2018. Accessed: June 5, 2019. [Online]. Available: www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#7e687cf60ba9

[4] M. Cox and D. Ellsworth, "Application-controlled demand paging for out-of-core visualization," in *Proc. Visualization'97 (Cat. No.97CB36155)*, 1997, pp. 235–244. doi: 10.1109/VISUAL.1997.663888.

[5] D. Laney, "3D data management: Controlling data volume, velocity, and variety," META Group, Westborough, MA, Tech. Rep., Feb. 2001. [Online]. Available: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

[6] D. Laney, "Deja VVVu: Others claiming Gartner's construct for big data," Gartner, Stamford, CT, Jan. 2012. [Online]. Available: https://blogs.gartner.com/doug-laney/deja-vvvue-others-claiming-gartners-volume-velocity-variety-construct-for-big-data/

[7] T. C. Havens, J. C. Bezdek, C. Leckie, L. O. Hall, and M. Palaniswami, "Fuzzy c-means algorithms for very large data," *IEEE Trans. Fuzzy Syst.*, vol. 20, no. 6, pp. 1130–1146, 2012. doi: 10.1109/TFUZZ.2012.2201485.

[8] P. Barnaghi, A. Sheth, and C. Henson, "From data to actionable knowledge: Big data challenges in the web of things," *IEEE Intell. Syst.*, vol. 28, no. 6, pp. 6–11, 2013. doi: 10.1109/MIS.2013.142.

[9] M. A. Khan, M. F. Uddin, N. Gupta, "Seven v's of big data understanding big data to extract value," in *Proc. 2014 Zone 1 Conf. American Society Engineering Education*, pp. 1–5. doi: 10.1109/ASEEZone1.2014.6820689.

[10] J. C. Bezdek and R. J. Hathaway, "VAT: A tool for visual assessment of (cluster) tendency," in *Proc. 2002 Int. Joint Conf. Neural Networks IJCNN'02 (Cat. No. 02CH37290)*, vol. 3, pp. 2225–2230. doi: 10.1109/IJCNN.2002.1007487.

[11] T. C. Havens and J. C. Bezdek, "An efficient formulation of the improved visual assessment of cluster tendency (iVAT) algorithm," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 5, pp. 813–822, 2011. doi: 10.1109/TKDE.2011.33.

[12] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, vol. 6. Englewood Cliffs, NJ: Prentice Hall 1988.

[13] B. S. Everitt, *Graphical Techniques for Multivariate Data*. Amsterdam, The Netherlands: North Holland, 1978.

[14] J. W. Tukey, *Exploratory Data Analysis,* Limited Preliminary ed. Reading, MA: Addison-Wesley, 1970.

[15] W. S. Cleveland, *Visualizing Data*. New Jersey: Hobart Press, 1993.

[16] L. Wang, U. T. Nguyen, J. C. Bezdek, C. A. Leckie, and K. Ramamohanarao, "iVAT and aVAT: Enhanced visual analysis for cluster tendency assessment," in *Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining*, 2010, pp. 16–27. doi: 10.1007/978-3-642-13657-3_5.

[17] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *J. Cybern.*, vol. 3, no. 3, pp. 32–57, 1973. doi: 10.1080/01969727308546046.

[18] R. J. Hathaway, J. C. Bezdek, and J. M. Huband, "Scalable visual assessment of cluster tendency for large data sets," *Pattern Recognit.*, vol. 39, no. 7, pp. 1315–1324, 2006. doi: 10.1016/j.patcog.2006.02.011.

[19] M. E. Johnson, L. M. Moore, and D. Ylvisaker, "Minimax and maximin distance designs," *J. Statist. Plan. Inference*, vol. 26, no. 2, pp. 131–148, 1990. doi: 10.1016/0378-3758(90)90122-B.

[20] T. C. Havens, J. C. Bezdek, and M. Palaniswami, "Scalable single linkage hierarchical clustering for big data," in *Proc. 2013 IEEE 8th Int. Conf. Intelligent Sensors, Sensor Networks and Information Processing*, pp. 396–401. doi: 10.1109/ISSNIP.2013.6529823.

[21] D. Kumar, J. C. Bezdek, M. Palaniswami, S. Rajasegarar, C. Leckie, and T. C. Havens, "A hybrid approach to clustering in big data," *IEEE Trans. Cybern.*, vol. 46, no. 10, pp. 2372–2385, 2015. doi: 10.1109/TCYB.2015.2477416.

[22] D. Kumar, M. Palaniswami, S. Rajasegarar, C. Leckie, J. C. Bezdek, and T. C. Havens, "clusiVAT: A mixed visual/numerical clustering algorithm for big data," in *Proc. 2013 IEEE Int. Conf. Big Data*, pp. 112–117. doi: 10.1109/BigData.2013.6691561.

[23] D. Kumar, H. Wu, S. Rajasegarar, C. Leckie, S. Krishnaswamy, and M. Palaniswami, "Fast and scalable big data trajectory clustering for understanding urban mobility," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 11, pp. 3709–3722, 2018. doi: 10.1109/TITS.2018.2854775.

[24] D. Kumar et al., "Adaptive cluster tendency visualization and anomaly detection for streaming data," *ACM Trans. Knowl. Discov. Data*, vol. 11, no. 2, p. 24, 2016. doi: 10.1145/2997656.

[25] P. Rathore, D. Kumar, J. C. Bezdek, S. Rajasegarar, and M. Palaniswami, "A rapid hybrid clustering algorithm for large volumes of high dimensional data," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 4, pp. 641–654, 2018. doi: 10.1109/TKDE.2018.2842191.

[26] M. Palaniswami, A. S. Rao, and S. Bainbridge, "Real-time monitoring of the great barrier reef using internet of things with big data analytics," *ITU J., ICT Discov.*, vol. 1, no. 13, pp. 1–10, 2017.

[27] J. C. Bezdek, S. Rajasegarar, M. Moshtaghi, C. Leckie, M. Palaniswami, and T. C. Havens, "Anomaly detection in environmental monitoring networks [application notes]," *IEEE Comput. Intell. Mag.*, vol. 6, no. 2, pp. 52–58, 2011. doi: 10.1109/MCI.2011.940751.

[28] D. Kumar, H. Wu, Y. Lu, S. Krishnaswamy, and M. Palaniswami, "Understanding urban mobility via taxi trip clustering," in *Proc. 2016 17th IEEE Int. Conf. Mobile Data Management (MDM)*, vol. 1, pp. 318–324. doi: 10.1109/MDM.2016.54.

[29] D. Kumar, S. Rajasegarar, M. Palaniswami, X. Wang, and C. Leckie, "A scalable framework for clustering vehicle trajectories in a dense road network," in *Proc. 4th Int. Workshop Urban Computing (UrbComp), Held Conjunction 21th ACM SIGKDD*, 2015, pp. 1–9.

[30] P. Rathore, D. Kumar, S. Rajasegarar, M. Palaniswami, and J. C. Bezdek, "A scalable framework for trajectory prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 10, pp. 3860–3874, 2019. doi: 10.1109/TITS.2019.2899179.

[31] D. Kumar, J. C. Bezdek, S. Rajasegarar, C. Leckie, and M. Palaniswami, "A visual-numeric approach to clustering and anomaly detection for trajectory data," *Visual Comput.*, vol. 33, no. 3, pp. 265–281, 2017. doi: 10.1007/s00371-015-1192-x.

SMC